Lord's Paradox and Targeted Interventions: The Case of Special Education

Roderick Theobald

A dissertation

submitted in partial fulfillments of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Dr. Thomas Richardson, co-chair

Dr. Dan Goldhaber, co-chair

Dr. Jon Wakefield

Program Authorized to Offer Degree:

Department of Statistics

University of Washington

Abstract

Lord's Paradox and Targeted Interventions: The Case of Special Education

Roderick Theobald

Co-chairs of the Supervisory Committee:

Dr. Thomas Richardson

Department of Statistics

Dr. Dan Goldhaber

Center for Education Data and Research

Lord (1967) describes a hypothetical paradox in which two statisticians, analyzing the same dataset using different but defensible methods, come to very different conclusions about the effects of an intervention on student outcomes. I use graphical methods–including a new graphical framework called Single World Object Oriented Plates (SWOOPs)–and detailed, longitudinal data about all public school students in Washington State to investigate a real-life example of Lord's Paradox that arises in evaluating the impact of special education services on student performance. I then introduce an instrumental variables (IV) approach that exploits a threshold in the state's special education funding laws that caps per-pupil special education funding at 12.7% of a districts students, and use SWOOPs to argue that the assumptions that justify this approach are more plausible than the assumptions that justify the methods in the existing literature. I find that students in districts that pass this threshold are far less likely to be placed in special education, all else equal, and use a district's position relative to this funding threshold as an instrumental variable to estimate the local average treatment effect of special education services on student test performance.

Contents

List of Figures

List of Tables

Preface

It seems that there are two primary approaches to research in applied statistics: some statisticians develop a method for an abstract problem and then seek out an application, while others start with an application and then try to improve upon prior approaches to the problem. This dissertation is squarely in the latter category. Since my time as a middle school math teacher in Oakland, CA, I have been interested in the impact of special education – a federally-funded program that provides support services to all students with diagnosed disabilities in U.S. public schools – on student outcomes. This question is substantively interesting to me because, at the outset of this project, I did not have a clear hypothesis. On the one hand, I was continually impressed by the dedication and persistence of friends and colleagues in Oakland who provided these services to students with disabilities. On the other hand, I was concerned that many teachers (myself included) held some students to a different standard of performance precisely *because* they were receiving special education services.

My first opportunity to explore the empirical evidence about this question was in a Causal Modeling course (taught by my co-advisor Thomas Richardson) three years ago. I was immediately struck by two observations that further motivated me to pursue this topic: (1) despite the fact that special education is the most-extensive and best-funded educational intervention in U.S. history, there is shockingly little large-scale, quantitative evidence about its impact on student performance; and (2) the research that does exist (e.g. Hanushek et al., 2002; Morgan et al., 2010) comes to contradictory conclusions. The latter observation is perhaps not as surprising as the former, because there are a number of potential explanations for the disparate findings in the existing literature on special education effects; specifically, these papers apply different methods to datasets from different time periods, locations, and grade levels. Fortunately, thanks to my affiliation with the Center for Education Data and Research (and my other co-advisor, CEDR director Dan Goldhaber, who gave me permission to use CEDR data for this project), I was able to apply the methods from Hanushek et al. (2002) and Morgan et al. (2010) to detailed, longitudinal data on public school students in Washington State. In this replication study – the final version of which is in chapter 3 of this dissertation – I found that, even when these methods were applied to the *exact same dataset*, one method suggested that special education has a *positive* and statistically-significant impact on student performance, while the other method suggested that the effect of special education is actually *negative* and statistically-significant.

This dissertation represents my attempt to understand this counterintuitive result. In my

ongoing work with Thomas Richardson, we have identified this phenomenon as a particularly striking example of "Lord's Paradox." Lord (1967) describes a hypothetical scenario in which two statisticians, analyzing the same dataset using different but defensible methods, come to very different conclusions about the effects of an intervention on student outcomes. In chapter 1, we introduce a different hypothetical example that illustrates how Lord's Paradox can arise when an educational intervention is targeted to low-performing students. We then use Single World Intervention Templates (Richardson and Robins, 2013), a recent unification of the potential outcomes and graphical approaches to causality, to describe and analyze three plausible scenarios: one in which statistician #1 is correct, one in which statistician #2 is correct, and one in which neither statistician is correct. We illustrate our conclusions with a simulation study, and apply the broad conclusions from this investigation to a highly-simplified model of special education placement in public schools.

Our work in chapter 1 illustrates the utility of graphical methods, and SWITs in particular, as a means of relating substantive scenarios to the counterfactual conditions that justify various statistical methods. Unfortunately, existing graphical methods proved to be insufficient when I tried to develop a more nuanced graphical discussion about special education. Specifically, the sheer number of variables (observed and otherwise) that I could relate to special education and student performance (and thus would have to be represented on a graph representing assumptions about special education placement in public schools) became too large. Motivated by this observation – and by the broader hypothesis that graphical methods are rarely used in education policy research for precisely this reason – we develop a new graphical framework in chapter 2 called Single World Object Oriented Plates (SWOOPs). SWOOPs are an extension of Directed Acyclic Graphs (DAGs), and facilitate causal reasoning in the multivariate, multilevel, and longitudinal settings that are common in empirical research in education. SWOOPs can be derived from DAGs in three steps: by aggregating related variables into "objects" (e.g. Koller and Pfeffer, 1997); by arranging objects within a level of data into "plates" (Buntine, 1994); and, following Richardson and Robins (2013), by introducing a "node-splitting" operation that allows for the inclusion of potential outcomes on the graph. We prove that conditional independence relationships in SWOOPs imply conditional independence relationships in any underlying DAG, and demonstrate how SWOOPs can be used to communicate and verify the assumptions that justify causal conclusions from observational data. As an application, we discuss value-added models (VAMs) of teacher effectiveness, and show that SWOOPs can be used to connect the substantive conditions (Rothstein, 2009) and the counterfactual conditions (Reardon and Raudenbush, 2009) that justify causal conclusions from different VAMs.

Having developed this new graphical methodology, I then return to the motivating question about special education in chapter 3. Specifically, I use data from Washington State to replicate the methods from Hanushek et al. (2002) and Morgan et al. (2010), and then use SWOOPs to discuss the counterfactual assumptions and substantive conditions that justify these methods. I conclude that the estimates from each study may be biased, but in opposite directions. I then introduce an instrumental variables (IV) approach that exploits a threshold in the state's special education funding laws that caps per-pupil special education funding at 12.7% of a district's students, and use SWOOPs to argue that the assumptions

that justify this approach are more plausible than the assumptions that justify the methods from Hanushek et al. (2002) and Morgan et al. (2010). I find that students in districts that pass this threshold are far less likely to be placed in special education, all else equal, and use a district's position relative to this funding threshold as an instrumental variable to estimate the average treatment effect of special education services on student test performance. With these data and instrument, I do not find evidence that, on average, special education services have a statistically significant impact on student performance.

In some ways, concluding this three-year odyssey without a definitive answer to the original, motivating question is unsatisfying. But I would rather have imprecise estimates I believe than precise estimates that I do not. After all, the precision of an estimate is partially a function of sample size, and the sample of students in Washington's longitudinal data system will only grow over time. I hold hope that, in my future work with these data, I can build on this work and provide more definitive evidence about this important topic.

Chapter 1

Lord's Paradox and Targeted Interventions

1.1 Introduction

1.1.1 Lord's Paradox

Lord (1967) describes a perplexing scenario that has come to be known as "Lord's Paradox". A university hires two statisticians to investigate the effects of the campus diet on student weights and any differences in these effects by gender. The first statistician compares the difference in average weight gains by gender and finds that the campus diet does not have a differential effect on the weights of boys and girls, while the second statistician uses analysis of covariance and finds that–controlling for differences in initial weights between the groups–boys show significantly more weight gain than girls. Lord concludes that "this paradox seems to impose a difficult interpretive task on those who wish to make similar studies of preformed groups."

Though Lord does little to resolve his own paradox, Holland and Rubin (1983) and Wainer (1991) use a potential outcomes framework (Neyman, 1923; Rubin, 1974) to demonstrate that each statistician's conclusion can be correct under different assumptions. A key insight in these discussions is that there is no control group in Lord's example; that is, there is no group that did not receive the campus diet. However, other authors (Jamieson, 1999; Maxwell and Delaney, 2004; Pearl, 2014; Wright, 2006) have noted the connection between Lord's paradox and pre/post studies in which there is a treatment group that receives an intervention and a control group that does not, but the treatment and control groups have very different average baseline measurements. One setting in which this can occur–discussed by Lord (1975), Holland and Rubin (1983), and Rubin et al. (2004)–is a "targeted intervention" in which individuals with low baseline measurements are more likely to receive the intervention. In section 1.1.2, we introduce a hypothetical example that illustrates how Lord's Paradox can arise in a targeted intervention.

1.1.2 A Tale of Two Statisticians

We modify Lord's original example as follows (though also see Lord (1975) for another school example). Suppose that a school develops a new reading intervention, and the school's principal receives funding for a "trial study" to investigate whether students score higher on a reading test after receiving the intervention than they would have in the absence of the intervention. Though the principal is interested in answering this long-term causal research question, she also views the funds for the trial study as a potentially valuable resource that can help raise the reading performance of low-performing students in the short term. She consequently works with teachers to identify students who need extra help in reading. These students (the "treatment group") receive the intervention, while other students in the school (the "control group") do not.

To evaluate the intervention, the school collects reading test scores for every student in the school in year 1, when none of the students received the intervention, and at the end of year 2, after students in the treatment group have received the intervention. Define the following observed variables for each student:

Definition 1.1.1 *A_2 = 1 if student is in treatment group in year 2 and 0 otherwise*

Definition 1.1.2 *Y_t = student's score on the test at the end of year t*

Suppose that the principal observes the data shown in figure 1.1.1 for the treatment ($A_2 = 1$) and control ($A_2 = 0$) groups.

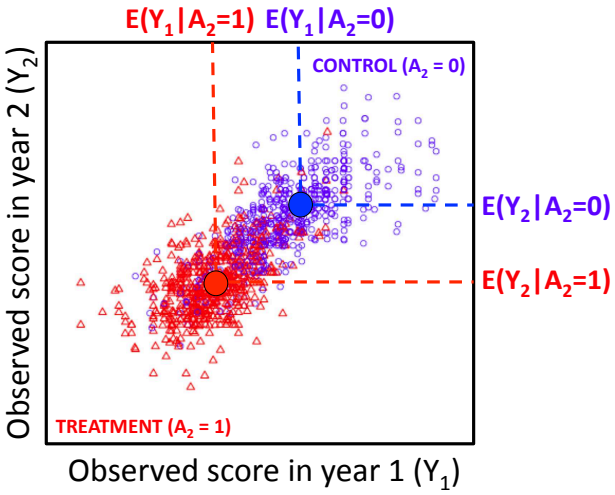


Figure 1.1.1: Hypothetical test score data from targeted intervention

Specifically, as in Lord's example, suppose that the observed joint distributions have the following three properties:

Property 1.1.1 *$E(Y_1|A_2 = 1) = E(Y_2|A_2 = 1) < E(Y_1|A_2 = 0) = E(Y_2|A_2 = 0)$ (i.e., the average scores within each group are the same in years 1 and 2, and the average scores in the treatment group are lower in both years than the average scores in the control group).*

Property 1.1.2 *$Var(Y_1|A_2 = 1) = Var(Y_2|A_2 = 1) = Var(Y_1|A_2 = 0) = Var(Y_2|A_2 = 0)$ (i.e., the variances of the scores are equal for each combination of group and year).*

Property 1.1.3 *$Corr(Y_1, Y_2|A_2 = 1) = Corr(Y_1, Y_2|A_2 = 0) < 1$ (i.e., the correlations between scores in year 1 and year 2 for each group are equal, and scores are not perfectly correlated between years).*

The principal recognizes that, since the treatment group has lower average baseline (pre-intervention) performance than the control group, it is not appropriate to evaluate the effect of the intervention by simply comparing the mean of Y_2 for each group. So, she hires two statisticians to evaluate the intervention by "adjusting" for the baseline differences in performance between the two groups.

Statistician #1 decides to adjust for baseline differences in performance by comparing the average "change" in test scores for the treatment and control groups: $\mathrm{E}(Y_2 - Y_1|A_2 = 1)$ and $\mathrm{E}(Y_2 - Y_1|A_2 = 0)$. The difference between these quantities is the well-known "differences-in-differences" estimator discussed in Angrist and Pischke (2008), for example. Statistician #1 calculates that the average change in each group is zero, since $\mathrm{E}(Y_2 - Y_1|A_2 = j) = \mathrm{E}(Y_2|A_2 = j) - \mathrm{E}(Y_1|A_2 = j) = 0$ for $j \in 0, 1$. This can be seen graphically in figure 1.1.2a, as the means of each group fall on the 45-degree line on which the value of Y_1 equals the value of Y_2. Statistician #1 concludes that the intervention has no average effect on student reading performance.

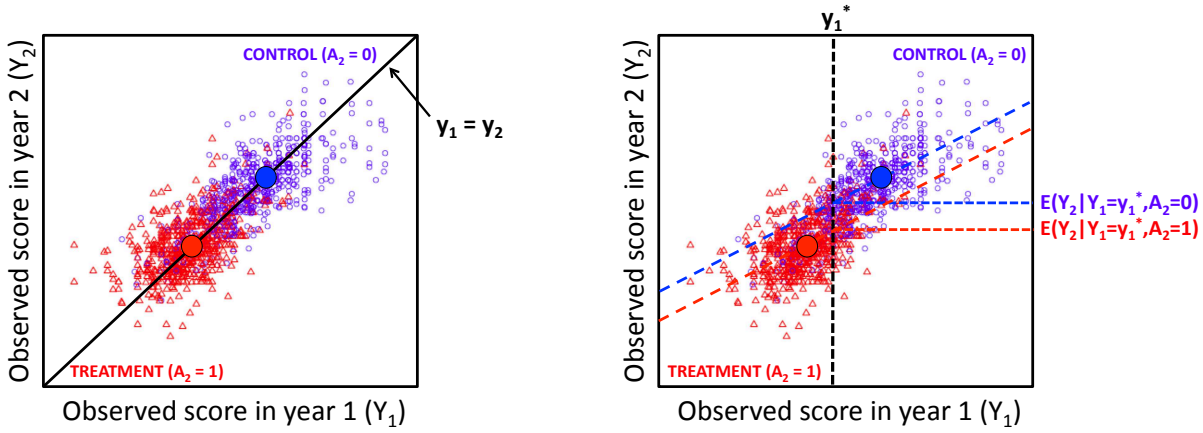
Statistician #2, on the other hand, decides to adjust for baseline differences in performance by estimating a linear regression model:

$$\mathrm{E}(Y_2|Y_1 = y_1, A_2 = a_2) = \beta_2 + \beta_{Y_1} y_1 + \beta_{A_2} a_2. \tag{1.1.1}$$

In equation 1.1.1, β_{Y_1} is the expected increase in Y_2 for each unit increase in the value of the baseline test score Y_1 (the slope of each regression line), while β_{A_2} is the average difference in the expected value of Y_2 between students with the value of $Y_1 = y_1$ in the treatment group ($A_2 = 1$) versus students with $Y_1 = y_1$ in the control group ($A_2 = 0$). Statistician #2 uses the least-squares estimate of β_{A_2} from equation 1.1.1 as an estimate of the causal effect of the intervention.

The diagonal lines in figure 1.1.2b show the estimated regression line for each group. For each value of Y_1 (including the value y_1^* shown in figure 1.1.2b), students in the treatment group are predicted to have a much lower value of Y_2; i.e., $\hat{\beta}_{A_2}$ is large and negative. From this, statistician #2 concludes that the intervention has a large negative effect on student reading performance.

So, which statistician is correct? Could each be correct under different assumptions? Or could both statisticians be incorrect?



(a) Figure for statistician #1

(b) Figure for statistician #2

Figure 1.1.2: Figures illustrating two approaches to the same dataset.

1.1.3 Resolving the Paradox

In this chapter, we describe three different substantive scenarios: one in which statistician #1 is correct; one in which statistician #2 is correct; and one in which neither statistician is correct. These substantive scenarios can be summarized as follows:

Scenario #1: Statistician #1 is correct when the intervention is targeted to students *solely* on the basis of student characteristics that are **time-invariant** (i.e., do not change over time) and **time-constant** in their effects on test scores in each year (i.e., have the same or no additive effect on student test scores in each year: see Wooldridge (2012)).

Scenario #2: Statistician #2 is correct when the intervention is targeted to students *solely* on the basis of **the student's test score in year 1** and the statistician's regression model is correctly specified (i.e., linearity holds and there are no interactions).

Scenario #3: Both statisticians are incorrect if the intervention is targeted to students on the basis of any **time-variant** characteristics that influence student test scores, or on the basis of factors that are **not time-constant** in their effects on test scores in each year.

We derive these conclusions in two steps. First, we introduce potential outcomes in section 1.2.1 and establish the counterfactual conditions that justify each statistician's approach. Second, after an introduction to graphical theory in section 1.2.2 and Single World Intervention Templates (SWITs) in section 1.2.3, we use SWITs in section 1.3 to describe and analyze the three scenarios above. This discussion is intended to illustrate the utility of graphical methods, and SWITs in particular, as a means of relating substantive scenarios to the counterfactual conditions that justify various statistical methods.

We also illustrate our primary conclusions in two ways. First, we present an easily replicable simulation in section 1.4 that illustrates each scenario and produces data that are consistent with the example in section 1.1.2. We then apply the conclusions from section 1.3 to a real-life example of Lord's Paradox in section 1.5 using data on the placement of public school students in Washington State into special education, the largest-scale "targeted intervention" in U.S. public schools. Finally, we offer some concluding thoughts in section 1.6.

1.2 Causality, Confounding, and Graphs

In our hypothetical example, the school's principal wants to know whether students *would* score higher on a reading test after receiving a reading intervention than they *would* score in the absence of the intervention. This type of question is central to a branch of statistics known as causal modeling. There are at least two prominent approaches to causality: potential outcomes (Neyman, 1923; Rubin, 1974) and directed acyclic graphs (DAGs) (e.g. Pearl, 1995). Recently, Richardson and Robins (2013) introduced a new unification of these approaches, Single World Intervention Templates (or SWITs). We briefly discuss these frameworks in the context of our hypothetical example.

1.2.1 Potential Outcomes

Overview and assumptions

The potential outcomes framework, introduced by Neyman (1923) and popularized by Rubin (1974), assumes the existence of variables describing the outcomes that would *potentially* be observed for each subject under a given treatment. In year 2 of the example, we can define two potential outcomes for student i: $Y_{2i}(a_{2i} = 1)$ is the test score of student i in year 2 if (potentially contrary to fact) student i received the intervention; and $Y_{2i}(a_{2i} = 0)$ is the test score of student i in year 2 if (potentially contrary to fact) student i did not receive the intervention. The statisticians in our example only observe one of these potential outcomes for each student: $Y_{2i}(a_{2i} = 1)$ if the student is in the treatment group; and $Y_{2i}(a_{2i} = 0)$ if the student is in the control group.

The notation $Y_{2i}(a_{2i})$ implies that a student's response to the intervention is the same regardless of whether and which other students in the school receive the intervention. This assumption may be problematic, as a student's response to the intervention may well depend on whether his or her friends and classmates also received the intervention. That is, if we let $\mathbf{a_2} \equiv (a_{2i}, \mathbf{a_{2,-i}})$ be the vector of treatment assignments for all students, where $\mathbf{a_{2,-i}}$ is the vector with a_{2i} removed, we could define the potential outcome for student i as $Y_{2i}(a_{2i}, \mathbf{a_{2,-i}})$ (i.e., as depending on the treatment assignment of all the other students in the school). If there are n students in the school, this would result in 2^n potential outcomes for each student. However, the potential outcomes literature often invokes the following stable unit-treatment value assumption, or SUTVA (Rubin, 1986):

Assumption 1.2.1 ***(SUTVA)****:* $a_{2i} = a'_{2i} \Rightarrow Y_{2i}(a_{2i}, \mathbf{a_{2,-i}}) = Y_{2i}(a'_{2i}, \mathbf{a'_{2,-i}}) \; \forall \; \mathbf{a'_{2,-i}}, \; i.$

SUTVA states that under two different assignments of students to the treatment and control groups, student i's potential outcomes are the same if he or she is assigned to the same group (treatment or control), *regardless of how other students are assigned to the treatment and control groups*. Recently, researchers have worked to relax SUTVA in educational settings (Hong and Raudenbush, 2006). However, for the remainder of this discussion we will assume that SUTVA holds, and that every student's potential outcomes depend only on whether he or she received the intervention (so that $Y_{2i}(a_{2i})$ is well-defined). From this point on, therefore, we will simplify notation and refer to each student's potential outcomes as $Y_2(a_2 = j)$ (or more compactly $Y_2(j)$) for $j \in 0, 1$ (note that we have dropped both the vector of treatment assignments for other students, $\mathbf{a_{2,-i}}$, and the subscript i).

Another important assumption that we will invoke throughout is consistency.

Assumption 1.2.2 ***(Consistency)****:* $A_2 = 1 \Rightarrow Y_2(1) = Y_2$ *and* $A_2 = 0 \Rightarrow Y_2(0) = Y_2$.

In other words, a student's test score when he or she receives the intervention is the same as the student's potential outcome under that treatment assignment. This assumption allows us to relate the expected values of potential outcomes to the expected values of observed variables via the formula $\mathrm{E}(Y_2(a_2)|A_2 = a_2) = \mathrm{E}(Y_2|A_2 = a_2)$.

Average treatment effect and confounding

As described earlier, the principal in our example would like to estimate the average difference between how students would have performed had they received the intervention and how students would have performed had they not received the intervention. This would be appropriate if the principal is thinking about giving the intervention to all students next year, for example. The estimand that the principal wants to estimate is called the "average treatment effect" of the intervention A_2 on student test scores Y_2, which can be expressed in terms of potential outcomes as $\mathrm{E}(Y_2(1) - Y_2(0))$.

The central concern in estimating the treatment effect of A_2 on Y_2 in our example is confounding. The "fundamental problem of causal inference" (Holland, 1986) is that we cannot observe both $Y_2(1)$ and $Y_2(0)$ for any individual student, which makes it difficult in general to produce unbiased estimates of $\mathrm{E}(Y_2(1))$ and $\mathrm{E}(Y_2(0))$ over all students. Without additional assumptions, we can only produce unbiased estimates of $\mathrm{E}(Y_2(1)|A_2 = 1) = \mathrm{E}(Y_2|A_2 = 1)$ (the average outcome under treatment among those in the treatment group) and $\mathrm{E}(Y_2(0)|A_2 = 0) = \mathrm{E}(Y_2|A_2 = 0)$ (the average outcome under control among those in the control group).

However, suppose that the treatment is randomized (note that this is *not* the case in our example!) Then knowing whether or not a student received the intervention in year 2 (A_2) gives no information about how the student would have scored on the test at the end of year 2, whether or not the student received ($Y_2(1)$) or did not receive ($Y_2(0)$) the intervention.

In other words, the treatment indicator for an individual student is independent of the potential outcomes for that student: $A_2 \perp\!\!\!\perp Y_2(1)$ and $A_2 \perp\!\!\!\perp Y_2(0)$. If these independences hold, then $\mathrm{E}(Y_2|A_2 = 1) = \mathrm{E}(Y_2(1)|A_2 = 1) = \mathrm{E}(Y_2(1))$ and $\mathrm{E}(Y_2|A_2 = 0) = \mathrm{E}(Y_2(0)|A_2 = 0) = \mathrm{E}(Y_2(0))$, so $\mathrm{E}(Y_2|A_2 = 1) - \mathrm{E}(Y_2|A_2 = 0) = \mathrm{E}(Y_2(1)|A_2 = 1) - \mathrm{E}(Y_2(0)|A_2 = 0) = \mathrm{E}(Y_2(1)) - \mathrm{E}(Y_2(0)) = \mathrm{E}(Y_2(1) - Y_2(0))$. That is, the average treatment effect is identified from the observed data. This leads to the following definition of marginal unconfounding (Rubin (1978) defines this as "ignorable" treatment assignment):

Definition 1.2.1 *The effect of A_2 on Y_2 is **marginally unconfounded** if $A_2 \perp\!\!\!\perp Y_2(a_2)\ \forall\ a_2$.*

But since the principal asked teachers to "target" the intervention to struggling students in our example, it should be clear that $A_2 \not\perp\!\!\!\perp Y_2(a_2)\ \forall\ a_2$ and that the treatment effect of A_2 on Y_2 is likely to be marginally confounded. Specifically, it is likely that the average of $Y_2(0)$ in the treatment group ($A_2 = 1$) is less than the average of $Y_2(0)$ in the control group ($A_2 = 0$), and that the average of $Y_2(1)$ in the treatment group is less than the average of $Y_2(1)$ in the control group. Or more intuitively, if the effect of A_2 on Y_2 were marginally unconfounded, there would be no reason to expect the differences in baseline performance between the treatment and control groups that exist in our example.

Each statistician in the example adjusts for the baseline differences in performance between the treatment and control groups in estimating the treatment effect of the intervention on student performance in year 2, but in different ways. Potential outcomes offer a way to describe the **counterfactual condition** that justifies each statistician's approach.

Counterfactual condition for statistician #1

Statistician #1 adjusts for baseline differences between the treatment and control groups by focusing on the "change score" $\Delta Y_2(a_2) \equiv Y_2(a_2) - Y_1$. Specifically, statistician #1 estimates the difference in the expected value of the observed change score ΔY_2 between students in the treatment group ($A_2 = 1$) and students in the control group ($A_2 = 0$):

$$\mathrm{E}(\Delta Y_2|A_2 = 1) - \mathrm{E}(\Delta Y_2|A_2 = 0). \tag{1.2.1}$$

This estimand equals the average treatment effect $\mathrm{E}(Y_2(1) - Y_2(0))$ if the effect of A_2 **on the change score ΔY_2** is marginally unconfounded:

Condition 1.2.1 *(Justifying statistician #1) $A_2 \perp\!\!\!\perp \Delta Y_2(a_2)\ \forall\ a_2$.*

The following derivation demonstrates why condition 1.2.1 implies that the quantity estimated by statistician #1 (equation 1.2.1) equals the average treatment effect $\mathrm{E}(Y_2(1) - Y_2(0))$:

$$\begin{aligned}
& \mathrm{E}(\Delta Y_2|A_2 = 1) - \mathrm{E}(\Delta Y_2|A_2 = 0) \\
= \ & \mathrm{E}(\Delta Y_2(1)|A_2 = 1) - \mathrm{E}(\Delta Y_2(0)|A_2 = 0) \text{ by assumption 1.2.2} \\
= \ & \mathrm{E}(\Delta Y_2(1)) - \mathrm{E}(\Delta Y_2(0)) \text{ by condition 1.2.1} \\
= \ & \mathrm{E}(Y_2(1) - Y_1) - \mathrm{E}(Y_2(0) - Y_1) \\
= \ & \mathrm{E}(Y_2(1)) - \mathrm{E}(Y_2(0)) + \mathrm{E}(Y_1) - \mathrm{E}(Y_1) \\
= \ & \mathrm{E}(Y_2(1) - Y_2(0)).
\end{aligned}$$

Note that in fact this derivation only requires that $\mathrm{E}(\Delta Y_2(a_2)|A_2 = a_2) = \mathrm{E}(\Delta Y_2(a_2)) \; \forall \; a_2$, which is weaker than condition 1.2.1. Nonetheless, we will focus on condition 1.2.1, as it is difficult to construct substantive scenarios in which $\mathrm{E}(\Delta Y_2(a_2)|A_2 = a_2) = \mathrm{E}(\Delta Y_2(a_2)) \; \forall \; a_2$ holds but condition 1.2.1 does not.

Counterfactual condition for statistician #2

Statistician #2 adjusts for baseline differences between the treatment and control groups by estimating a linear regression model (equation 1.1.1) that controls for each student's score in year 1. This approach is an example of covariate adjustment methods that estimate the average difference in the expected value of Y_2 between students with the value of $Y_1 = y_1$ in the treatment group ($A_2 = 1$) versus students with $Y_1 = y_1$ in the control group ($A_2 = 0$):

$$\mathrm{E}\{\mathrm{E}(Y_2|Y_1 = y_1, A_2 = 1) - \mathrm{E}(Y_2|Y_1 = y_1, A_2 = 0)\}. \tag{1.2.2}$$

This estimand equals the average treatment effect $\mathrm{E}(Y_2(1) - Y_2(0))$ if the effect of A_2 on Y_2 is unconfounded **conditional on the baseline test score $\mathbf{Y_1}$**:

Condition 1.2.2 *(Justifying statistician #2)* $A_2 \perp\!\!\!\perp Y_2(a_2) \mid Y_1 \; \forall \; a_2$.

The following derivation demonstrates why condition 1.2.2 implies that the quantity estimated by statistician #2 (equation 1.2.2) equals the average treatment effect $\mathrm{E}(Y_2(1) - Y_2(0))$:

$$\begin{aligned}
& \mathrm{E}\{\mathrm{E}(Y_2|Y_1 = y_1, A_2 = 1) - \mathrm{E}(Y_2|Y_1 = y_1, A_2 = 0)\} \\
= \; & \mathrm{E}\{\mathrm{E}(Y_2(1)|Y_1 = y_1, A_2 = 1) - \mathrm{E}(Y_2(0)|Y_1 = y_1, A_2 = 0)\} \text{ by assumption 1.2.2} \\
= \; & \mathrm{E}\{\mathrm{E}(Y_2(1)|Y_1 = y_1) - \mathrm{E}(Y_2(0)|Y_1 = y_1)\} \text{ by condition 1.2.2} \\
= \; & \mathrm{E}\{\mathrm{E}(Y_2(1) - Y_2(0)|Y_1 = y_1)\} \\
= \; & \mathrm{E}(Y_2(1) - Y_2(0))
\end{aligned}$$

Much as the counterfactual condition justifying the conclusion of statistician #1 is stronger than necessary, we note that the derivation above only requires that $\mathrm{E}(Y_2(a_2)|Y_1 = y_1, A_2 = a_2) = \mathrm{E}(Y_2(a_2)|Y_1 = y_1) \; \forall \; a_2$, which is weaker than condition 1.2.2. Nonetheless, we focus on condition 1.2.2 because it is difficult to construct substantive scenarios in which $\mathrm{E}(Y_2(a_2)|Y_1 = y_1, A_2 = a_2) = \mathrm{E}(Y_2(a_2)|Y_1 = y_1) \; \forall \; a_2$ but condition 1.2.2 does not hold. Condition 1.2.2 is sometimes called the "Conditional Independence Assumption" (Angrist and Pischke, 2008), and combined with the assumption that $0 < \Pr(A_2 = 1|Y_1 = y_1) < 1$, is equivalent to the oft-cited "conditional ignorability" condition of Rosenbaum and Rubin (1983).

We also note that the linear regression model estimated by statistician #2 (equation 1.1.1) imposes additional parametric assumptions; namely, that the relationship between Y_1 and Y_2 is linear, and that there is no interaction between the treatment and the value of Y_1. These assumptions are avoidable, however, as statistician #2 could have used a different covariate adjustment approach that allows for a non-linear (or non-parametric) relationship between Y_1 and Y_2, or allows for an interaction between Y_1 and the treatment. So, though we describe the approach of statistician #2 as linear regression to preserve the

analogy with Lord's original example, we focus on condition 1.2.2 in section 1.3 (when we describe substantive scenarios that justify each statistician's conclusion), as it is a common assumption of a broad class of covariate adjustment methods.

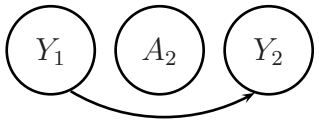
1.2.2 DAGs

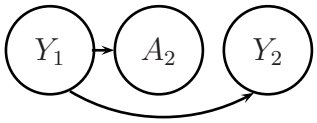
We now discuss our hypothetical example using graphs. Directed acyclic graphs (DAGs) (e.g. Pearl, 1995) are one useful type of graph, particularly for reasoning about joint distributions.

Definition 1.2.2 *A directed acyclic graph (DAG) is a graph containing directed edges ($\rightarrow$) and no directed cycles ($V \rightarrow \ldots \rightarrow V$).*

We will first use the observed variables in our example to give a general introduction to DAGs, and then introduce one type of unobserved variable and a condition called d-separation that will allow us to discuss and establish independence relationships between the observed variables.

In our example, each statistician observes three variables about each student: the test scores in each year, Y_1 and Y_2, and the treatment indicator A_2. Figure 1.2.1 contains two DAGs that illustrate potential relationships between these variables.





(a) $P(Y_2 = y_2, A_2 = a_2, Y_1 = y_1)$
$= P(Y_2 = y_2|Y_1 = y_1)P(A_2 = a_2)P(Y_1 = y_1)$

(b) $P(Y_2 = y_2, A_2 = a_2, Y_1 = y_1)$
$= P(Y_2 = y_2|Y_1 = y_1)P(A_2 = a_2|Y_1 = y_1)P(Y_1 = y_1)$

Figure 1.2.1: Two DAGs illustrating possible relationships between the observed variables.

Each DAG in figure 1.2.1 implies a different factorization of the joint distribution of Y_1, Y_2, and A_2. One way to infer the factorization of a distribution from a DAG is to focus on the "parents" of each vertex, defined as follows:

Definition 1.2.3 *The **parents** of a vertex V with respect to a graph $\mathcal{G}$ are $pa_{\mathcal{G}}(V) = \{X|X \rightarrow V\}$.*

For example, Y_1 is a parent of Y_2 in figure 1.2.1a, while Y_1 is also a parent of A_2 in figure 1.2.1b.

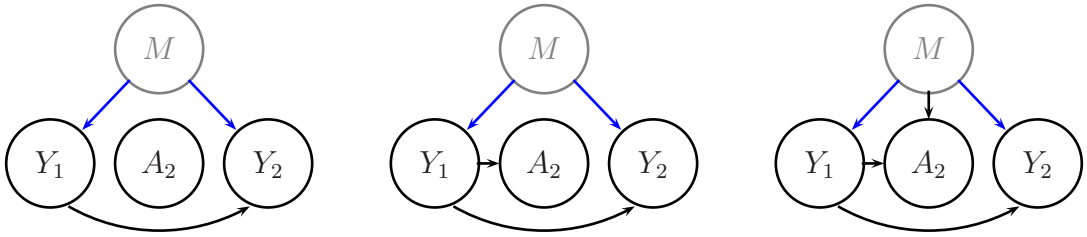
The joint distribution of Y_1, Y_2, and A_2 can always be factorized as $P(Y_2 = y_2, A_2 = a_2, Y_1 = y_1) = P(Y_2 = y_2|A_2 = a_2, Y_1 = y_1)P(A_2 = a_2|Y_1 = y_1)P(Y_1 = y_1)$. However, missing edges in a DAG imply conditional independence assumptions that we can use to simplify this factorization. We will assume throughout this paper that distributions are Markov with respect to the DAG.

Definition 1.2.4 *A distribution P is **Markov** with respect to a DAG $\mathcal{G}$ if* $P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V|pa_{\mathcal{G}}(V))$.

In other words, we will assume that the joint distribution of the variables in a DAG can be factorized as the product of the distributions of each variable conditional on its parents. The relationships shown in figure 1.2.1a, then, allow us to factorize the joint distribution of Y_1, Y_2, and A_2 as $P(Y_2 = y_2, A_2 = a_2, Y_1 = y_1) = P(Y_2 = y_2|Y_1 = y_1)P(A_2 = a_2)P(Y_1 = y_1)$. Likewise, the relationships shown in figure 1.2.1b allow us to factorize the joint distribution as $P(Y_2 = y_2, A_2 = a_2, Y_1 = y_1) = P(Y_2 = y_2|Y_1 = y_1)P(A_2 = a_2|Y_1 = y_1)P(Y_1 = y_1)$.

At various times in our discussion in sections 1.3.1-1.3.3, we will use graphs to establish independence relationships between variables. We will also occasionally assume linear, additive relationships between variables in a DAG (particularly in our simulations in section 1.4). To introduce these concepts, suppose there is an *unobserved* variable M that is *time-invariant* (i.e., does not change over time) and *time-constant* in its effect on test scores in each year (see equations 1.2.3 and 1.2.4). In the economics literature, M is often called student "ability", but in reality it can be any time-invariant student characteristic that is assumed to have a consistent impact on student test performance over time.

Figure 1.2.2 shows three possible sets of relationships between the three observed variables shown in figure 1.2.1 and the unobserved variable M (we refer to a variable as unobserved if it is unobserved *to the statisticians in our example*. The principal, on the other hand, may have additional information about M, such as an IQ score).



(a) $A_2 \perp\!\!\!\perp Y_2$ (b) $A_2 \perp\!\!\!\perp Y_2|Y_1$ (c) $A_2 \not\perp\!\!\!\perp Y_2|Y_1$

Figure 1.2.2: Three DAGs illustrating (a) independence; (b) conditional independence given Y_1; and (c) conditional dependence given Y_1 of A_2 and Y_2. See equations 1.2.3 and 1.2.4.

In figure 1.2.2 (and in all figures in this paper), black vertices represent observed variables while gray vertices represent unobserved variables. In each sub-figure, the blue edges from M to Y_1 and Y_2 represent the assumption that M has the *same* linear effect on Y_1 and Y_2. To make this concrete, we introduce the assumption that the edges into Y_1 and Y_2 in each DAG in figure 1.2.2 represent linear, additive relationships.

$$\mathrm{E}(Y_1|M = m, A_2 = a_2) = \beta_1 + \beta_M m \quad (1.2.3)$$

$$\mathrm{E}(Y_2|M = m, Y_1 = y_1, A_2 = a_2) = \beta_2 + \beta_M m + \beta_{Y_1} y_1. \quad (1.2.4)$$

The common coefficient β_M in equations 1.2.3 and 1.2.4 represents the linear, time-constant effect of M on test scores in each year. Note also that none of the DAGs in 1.2.2 have an edge between A_2 and Y_2 (and neither of the expected values in equations 1.2.3 and 1.2.4 depend on the value a_2 of A_2). This represents the null hypothesis that the intervention A_2 does not have an effect on student test scores Y_2. This will be relaxed in the next section.

The differences between the sub-figures in figure 1.2.2 concern the presence or absence of edges from Y_1 and M to A_2. In sub-figure 1.2.2a, the absence of these edges implies that the probability that a student is assigned to the treatment group is independent of the value of both Y_1 and A_2 for that student. The edge between Y_1 and A_2 in sub-figure 1.2.2b implies that assignment to the treatment group depends (at least in part) on Y_1. The additional edge between M and A_2 in sub-figure 1.2.2c implies that assignment to the treatment group also depends (at least in part) on M.

Establishing conditional independence relationships between variables will play a central role in our discussion. The condition in the DAG literature that establishes conditional independence is known as d-separation (Pearl, 2000). Before presenting this definition and the theorem that relates it to independence relationships, we provide some background definitions.

Definition 1.2.5 *A **path** π between vertices X and Y consists of a sequence of distinct vertices that are connected by edges.*

For example, there are two paths from M to Y_2 in figure 1.2.2a: $M \to Y_1 \to Y_2$ and $M \to Y_2$.

Definition 1.2.6 *A vertex A is an **ancestor** of a vertex D if one of two conditions holds: (a) there is a directed path $A \to \ldots \to D$ from A to D; or (b) $A = D$.*

For example, M is an ancestor of A_2 in figure 1.2.2b even though there is no edge between M and A_2 because of the directed path $M \to Y_1 \to A_2$.

Definition 1.2.7 *A non-endpoint vertex V is a **collider** on a path π if π takes the form $X \ldots \to V \leftarrow \ldots Y$. Non-endpoints on π that are not colliders are **non-colliders** on π.*

For example, A_2 is a collider on the path $M \to A_2 \leftarrow Y_1$ in figure 1.2.2c.

With these definitions, we can define d-separation (e.g. Pearl, 2000):

Definition 1.2.8 *A path π in a graph $\mathcal{G}$ **d-connects** vertices A and B conditional on a set $\mathbf{C}$ in $\mathcal{G}$ if A and B are the endpoints of π, every non-collider on π is not in $\mathbf{C}$, and every collider on π is an ancestor of $\mathbf{C}$ (or is in $\mathbf{C}$.) If there is no path d-connecting A and B given $\mathbf{C}$ in $\mathcal{G}$, then A and B are **d-separated** given $\mathbf{C}$ in $\mathcal{G}$.*

The primary application of d-separation is to establish conditional independence relationships, as formalized in the following well-known results (e.g. Lauritzen et al., 1990):

Theorem 1.2.1 *In any distribution P that is Markov with respect to $\mathcal{G}$, if A and B are d-separated given $\mathbf{C}$ in $\mathcal{G}$ then $A \perp\!\!\!\perp B|\mathbf{C}$ in P.*

Theorem 1.2.2 *If A and B are d-connected given $\mathbf{C}$ in $\mathcal{G}$, then there exists a (multivariate Gaussian) distribution that is Markov with respect to $\mathcal{G}$ in which $A \not\perp\!\!\!\perp B|\mathbf{C}$ in P.*

We can apply Theorems 1.2.1 and 1.2.2 to the DAGs in figure 1.2.2 and "work out" the conditional independence relationships implied by each graph. For example, in figure 1.2.2a, there is no path connecting A_2 and Y_2, and thus (trivially) no path that d-connects A_2 and Y_2 given the empty set. Thus A_2 and Y_2 are d-separated given the empty set in figure 1.2.2a and, by Theorem 1.2.1, $A_2 \perp\!\!\!\perp Y_2$ in any distribution P that is Markov with respect to the graph in figure 1.2.2a. From this point on, we will assume that the Markov property holds and simply say in this scenario that the relationships in figure 1.2.2a imply that $A_2 \perp\!\!\!\perp Y_2$.

The DAG in figure 1.2.2b contains the paths $A_2 \leftarrow Y_1 \rightarrow Y_2$ and $A_2 \leftarrow Y_1 \leftarrow M \rightarrow Y_2$. Y_1 is a non-collider on each path, so these paths d-connect A_2 and Y_2 given the empty set. Thus by Theorem 1.2.2, there exists a distribution P that is Markov with respect to the DAG in figure 1.2.2b such that $A_2 \not\perp\!\!\!\perp Y_2$. While it is still possible to construct distributions that are Markov with respect to these graphs yet $A_2 \perp\!\!\!\perp Y_2$, we will use the convention from here on that we will write $A_2 \not\perp\!\!\!\perp Y_2$ if there exists a distribution Markov with respect to the DAG in which there is dependence. Note that neither of the paths between A_2 and Y_2 d-connects A_2 and Y_2 given Y_1, since Y_1 is a non-collider on each path. Thus A_2 and Y_2 are d-separated given Y_1 and $A_2 \perp\!\!\!\perp Y_2|Y_1$ in figure 1.2.2b.

Figure 1.2.2c, on the other hand, contains the additional path $A_2 \leftarrow M \rightarrow Y_2$. The only non-collider on this path is M which is not in $\{Y_1\}$, and there are no colliders on the path. So, this path d-connects A_2 and Y_2 given Y_1. Thus A_2 and Y_2 are not d-separated given Y_1 and $A_2 \not\perp\!\!\!\perp Y_2|Y_1$ in figure 1.2.2c.

There is a direct connection between these conditional independence relationships and the coefficients in a linear regression model. Consider, for example, the linear regression model estimated by statistician #2 in our example:

$$\mathrm{E}(Y_2|Y_1 = y_1, A_2 = a_2) = \beta_2 + \beta_{Y_1} y_1 + \beta_{A_2} a_2 \tag{1.2.5}$$

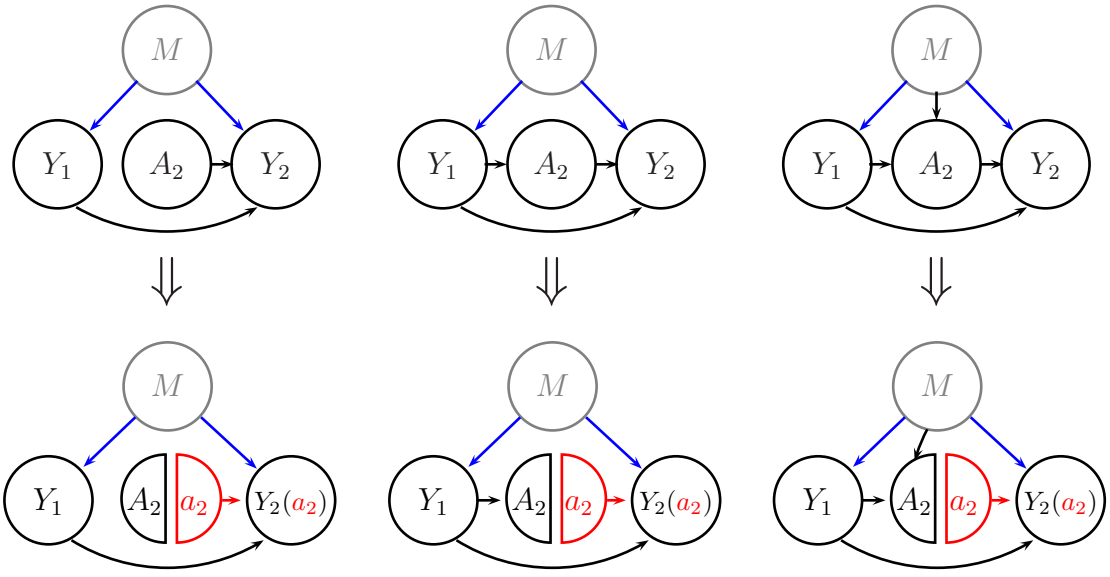
The relationships shown in figures 1.2.2a and 1.2.2b imply that $A_2 \perp\!\!\!\perp Y_2|Y_1$, which in turn implies that $\beta_{A_2} = 0$ in equation 1.2.5. The relationships shown in figure 1.2.2c, on the other hand, imply that $A_2 \not\perp\!\!\!\perp Y_2|Y_1$, which in turn implies that $\beta_{A_2} \neq 0$, even though there is no edge between A_2 and Y_2 in the DAG in figure 1.2.2c. This is a consequence of the confounding path $A_2 \leftarrow M \rightarrow Y_2$ in this DAG.

1.2.3 SWITs

We can now introduce the graphs that we will use in our discussion in section 1.3, known as Single World Intervention Templates (SWITs) (Richardson and Robins, 2013). SWITs are a unification of potential outcomes and DAGs, and offer a number of advantages over traditional DAGs. First, SWITs permit the inclusion of potential outcomes in a graphical

framework, which will allow us to connect the conditional independence results from the DAG literature (theorems 1.2.1 and 1.2.2) to the potential outcomes definition of confounding (definition 1.2.1). Second, SWITs distinguish between variables that are manipulable and variables that are not. For example, it is (arguably) not possible to manipulate a student's ability (M) directly, but it is certainly possible to manipulate whether or not a student receives the intervention (A_2). DAGs do not distinguish between these types of variables, while SWITs do.

To illustrate the difference between DAGs and SWITs, figure 1.2.3 shows SWITs that are derived from the DAGs in figure 1.2.2, except under the alternative hypothesis that the intervention A_2 *does* have an effect on the outcome Y_2. There are two key differences between the SWITs in figure 1.2.3 and the DAGs from which they are derived. First, the vertices that represent the intervention in the DAGs have been "split" into the random variable A_2 and the hypothetical intervention a_2. This indicates that this is the variable that is being manipulated in our example. Second, the vertices that represent the outcome Y_2 in the DAGs have been relabeled as potential outcomes that depend on the hypothetical intervention a_2. This allows us to apply the results from the DAG literature (theorems 1.2.1 and 1.2.2) to potential outcomes.



(a) $A_2 \perp\!\!\!\perp Y_2(a_2)\ \forall\ a_2$ (b) $A_2 \perp\!\!\!\perp Y_2(a_2)|Y_1\ \forall\ a_2$ (c) $A_2 \not\perp\!\!\!\perp Y_2(a_2)|Y_1\ \forall\ a_2$

Figure 1.2.3: Single World Intervention Templates (SWITs) derived from Directed Acyclic Graphs (DAGs)

Each of the SWITs in figure 1.2.3 actually represents two graphs (Single World Intervention Graphs, or SWIGs) for each student: one in a hypothetical world in which the student received the intervention ($a_2 = 1$), and the other in a hypothetical world in which the student did not receive the intervention ($a_2 = 0$). We will assume an additive model throughout (see equations 1.3.1 and 1.3.2), meaning that the relationships between variables do not change depending on the value a_2 (i.e., there is no interaction term in equations 1.3.1 and 1.3.2).

Because of this, any conclusions we can draw from a SWIT are true for all values of the intervention (i.e., $\forall\ a_2$). See Richardson and Robins (2013) for more details about SWIGs.

Just as missing edges in a DAG imply conditional independence assumptions, missing edges in a SWIT imply conditional independencies that allow us to simplify joint distributions that include potential outcomes. As before, we assume that the Markov property (definition 1.2.4) holds and the variables in a SWIT can be factorized as the product of the distributions of each variable conditional on its parents. Then the relationships shown in the SWIT in figure 1.2.3a, for example, allow us to factorize the joint distribution of Y_1, M, A_2, and $Y_2(a_2)$, as $P(Y_2(a_2) = y_2, A_2 = a_2^*, M = m, Y_1 = y_1) = P(Y_2(a_2) = y_2|Y_1 = y_1, M = m, A_2 = a_2)P(Y_1 = y_1|M = m)P(M = m)P(A_2 = a_2^*)$.

SWITs also provide a direct connection between the d-separation results from the DAG literature (theorems 1.2.1 and 1.2.2) and the conditional independence relationships in the potential outcomes definition of marginal unconfounding (definition 1.2.1). For example, in the SWIT in figure 1.2.3a, there is no path that d-connects A_2 and $Y_2(a_2)$ given the empty set; that is, $A_2 \perp\!\!\!\perp Y_2(a_2)\ \forall\ a_2$. Thus directly from definition 1.2.1, we can conclude that the effect of A_2 on Y_2 is marginally *unconfounded*. But as we saw in section 1.2.2, the path $A_2 \leftarrow Y_1 \rightarrow Y_2$ in the SWITs in figures 1.2.3b and 1.2.3c d-connects A_2 and $Y_2(a_2)$ given the empty set, so $A_2 \not\perp\!\!\!\perp Y_2(a_2)\ \forall\ a_2$. Thus the relationships in these SWITs imply that the effect of A_2 on Y_2 is marginally *confounded*.

SWITs can further be used to establish whether the effect of A_2 on Y_2 is conditionally unconfounded (condition 1.2.2). By the exact same d-separation argument we applied to figure 1.2.2b, we can demonstrate that $A_2 \perp\!\!\!\perp Y_2(a_2)|Y_1$ in the SWIT in figure 1.2.3b. This is precisely condition 1.2.2, so we can conclude that the effect of A_2 on Y_2 is *unconfounded* conditional on Y_1 in figure 1.2.3b. On the other hand, we can use the same d-separation argument we applied to the DAG in figure 1.2.2c to the SWIT in figure 1.2.3c and show that $A_2 \not\perp\!\!\!\perp Y_2(a_2)|Y_1$. Thus we can conclude that the effect of A_2 on Y_2 is *confounded*, even conditional on Y_1, in figure 1.2.3c.

1.3 Application to Lord's Paradox

We can now return to our central question: under what substantive scenarios does each statistician produce an unbiased estimate of the causal effect of the intervention on student performance? To describe these scenarios, we introduce a slightly more complicated (and realistic) SWIT than those in figure 1.2.3. The central problem with the relationships shown in figure 1.2.3 is that, although Y_1 and Y_2 are certainly correlated, Y_1 does not truly have a "causal effect" on Y_2. Instead, the correlation between these test scores arises from two sources. The first is already represented in figure 1.2.3: the effect of *time-invariant* and *time-constant* characteristics M on Y_1 and Y_2. The second, though, is the effect of any characteristics that change over time or do not have the same effect on test scores in each year (and are not influenced by test scores in either year). These characteristics can be the

quality of the student's teacher and school in each year, parental influence in each year, and the student's motivation level in each year. We add these time-variant characteristics to figure 1.3.1 as U_1 and U_2.

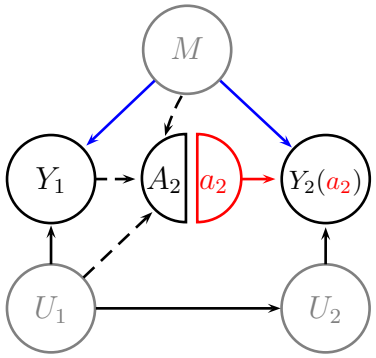


Figure 1.3.1: SWIT for motivating example. See equations 1.3.1 and 1.3.2.

In figure 1.3.1, time-variant characteristics in year t, U_t, have an effect on test scores in year t, Y_t. The edge between U_1 and U_2 indicates that these characteristics are correlated between years (e.g., a student's motivation level in year 1 is likely to be correlated with his or her motivation level in year 2). Notice that we place no assumptions on the relationships between these edges (e.g., the effect of U_1 on Y_1 can be different than the effect of U_2 on $Y_2(a_2)$), and that we do not include any other potential confounders in figure 1.3.1. This is partially to preserve the analogy with Lord's original example (Lord, 1967), but also because it simply is not feasible to include all the potential confounders from this example in a SWIT. This motivates our development, in chapter 2, of a graphical framework that can accommodate any number of confounders, observed or otherwise.

We can now modify equations 1.2.3 and 1.2.4 to reflect the relationships in figure 1.3.1.

$$\begin{aligned} \mathrm{E}(Y_1|M=m,\mathbf{U}=\mathbf{u}) &= \beta_1+\beta_M m+\beta_{U_1}u_1 && (1.3.1)\\ \mathrm{E}(Y_2(a_2)|M=m,\mathbf{U}=\mathbf{u},A_2=a_2) &= \beta_2+\beta_M m+\beta_{U_2}u_2+\beta_{A_2}a_2. && (1.3.2)\end{aligned}$$

As before, the edges into Y_1 in figure 1.3.1 represent linear, additive relationships, but Y_1 is now a function of the values of both M and U_1 (equation 1.3.1). Equation 1.3.2 actually represents two different equations: one for $a_2 = 1$ (i.e., in a hypothetical world when the student receives the intervention), and the other for $a_2 = 0$ (i.e., in a hypothetical world when the student does not receive the intervention). β_{A_2} in equation 1.3.2 represents the edge from a_2 to $Y_2(a_2)$ in figure 1.3.1. We stress that β_{A_2} may be different for different students (i.e., the intervention may impact different students differently.) However, we can still define the average treatment effect as $\beta_{ATE} \equiv \mathrm{E}(\beta_{A_2})$. Notice that equation 1.3.2 implies

that $\beta_{ATE} = \mathrm{E}(Y_2(1) - Y_2(0))$, the definition of average treatment effect from section 1.2.1:

$$\begin{aligned}
\mathrm{E}(Y_2(1) - Y_2(0)) &= \mathrm{E}(Y_2(1)) - \mathrm{E}(Y_2(0)) \\
&= \mathrm{E}\{\mathrm{E}(Y_2(1)|M = m, U_2 = u_2)\} - \mathrm{E}\{\mathrm{E}(Y_2(0)|M = m, U_2 = u_2)\} \\
&= \mathrm{E}\{\mathrm{E}(Y_2(1)|M = m, U_2 = u_2, A_2 = 1)\} - \mathrm{E}\{\mathrm{E}(Y_2(0)|M = m, U_2 = u_2, A_2 = 0)\} \\
&= \mathrm{E}(\beta_2 + \beta_M m + \beta_{U_2} u_2 + \beta_{A_2}) - \mathrm{E}(\beta_2 + \beta_M m + \beta_{U_2} u_2) \\
&= \mathrm{E}(\beta_{A_2}) \equiv \beta_{ATE}
\end{aligned}$$

Equations 1.3.1 and 1.3.2 will form the basis of our simulations in section 1.4, and represent the assumptions we will make throughout this section. We will also assume throughout that the distributions of $Y_1|M, \mathbf{U}$ and $Y_2(a_2)M, \mathbf{U}, A_2$ are conditional Gaussian with constant variance so that, for example, the absence of u_1 in equation 1.3.2 implies:

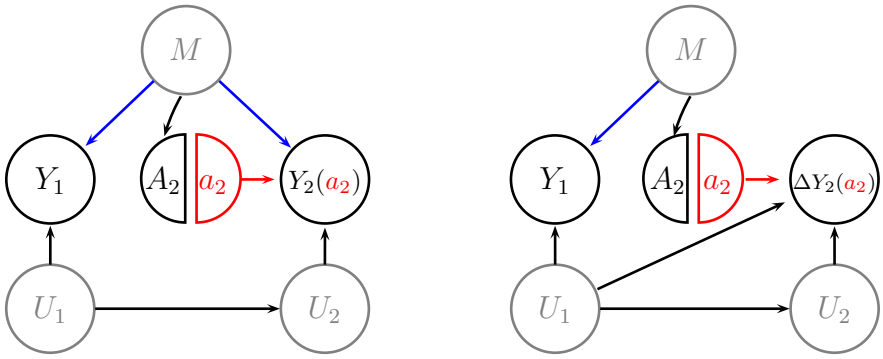
$$U_1 \perp\!\!\!\perp Y_2(a_2)|M, U_2, A_2. \tag{1.3.3}$$

We now describe substantive scenarios in sections 1.3.1 and 1.3.2 under which each statistician is correct, and also describe a scenario in section 1.3.3 under which neither statistician is correct. Each substantive scenario can be expressed graphically in terms of the presence or absence of the three dashed edges in figure 1.3.1. In a randomized control trial, none of these edges would exist (i.e., no student characteristics would be correlated with the probability of receiving the intervention). But given that the intervention in the motivating example is targeted to students who are struggling in reading, the key question is *how* students are being targeted. Are they being targeted based on their test score in year 1? If so, the edge between Y_1 and A_2 exists. Are they being targeted based on a measure of some time-invariant, time-constant characteristic, like an IQ score? If so, the edge between M and A_2 exists. Or are they being targeted based on unobserved time-variant characteristics, such as their attitude, motivation level, or parental involvement in year 1? If so, the edge between U_1 and A_2 exists.

1.3.1 When is Statistician #1 Correct?

Suppose that a student's ability, as measured by his or her IQ score, is a time-invariant and time-constant student characteristic (as is often assumed in the economics literature). To identify "low-performing students", the principal in our example might choose to administer an IQ test to all students and assign all students in the lower half of the distribution of IQ scores to the treatment group (students in the upper half of the distribution then become the control group). Figure 1.3.2a shows the causal relationships between unobserved and observed variables in this scenario.

The key difference between figure 1.3.2a and figure 1.3.1 is the *presence* of the edge between M and A_2 and the *absence* of the edges between Y_1 and A_2 and between U_1 and A_2. The edge between M and A_2 implies that students are selected for the intervention on the basis of a time-invariant, time-constant characteristic: in this case, the student's IQ score. The absence of the edges between Y_1 and A_2 and between U_1 and A_2 implies that students are *not* selected for the intervention either on the basis of prior year test scores or on *any*



(a) SWIT for $Y_2(a_2)$ (b) SWIT for $\Delta Y_2(a_2)$

Figure 1.3.2: SWITs under substantive scenario #1

other characteristic that changes over time.

Our goal is to show that the relationships in figure 1.3.2a imply condition 1.2.1 ($A_2 \perp\!\!\!\perp \Delta Y_2(a_2) \; \forall \; a_2$), the condition that justifies the conclusion of statistician #1. As a preliminary step, notice that the only path between Y_1 and A_2 in figure 1.3.2a is $Y_1 \leftarrow M \rightarrow A_2$. M is a non-collider on this path, so by definition 1.2.8, Y_1 is d-separated from A_2 given $\{M, \mathbf{U}\}$. Under the assumption that the distribution P is Markov with respect to the SWIT in figure 1.3.2a, theorem 1.2.1 implies that $Y_1 \perp\!\!\!\perp A_2 | M, \mathbf{U}$. This allows us to write:

$$\mathrm{E}(Y_1 | M = m, \mathbf{U} = \mathbf{u}) \quad = \quad \mathrm{E}(Y_1 | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2). \tag{1.3.4}$$

The next step, then, is to replace the $Y_2(a_2)$ vertex in figure 1.3.2a with a vertex for the change score $\Delta Y_2(a_2) \equiv Y_2(a_2) - Y_1$, shown in figure 1.3.2b. We can derive the edges into this new vertex from equations 1.3.1 and 1.3.2.

$$\begin{aligned}
& \mathrm{E}(\Delta Y_2(a_2) | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2) \\
= \; & \mathrm{E}(Y_2(a_2) - Y_1 | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2) \\
= \; & \mathrm{E}(Y_2(a_2) | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2) - \mathrm{E}(Y_1 | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2) \\
= \; & \mathrm{E}(Y_2(a_2) | M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2) - \mathrm{E}(Y_1 | M = m, \mathbf{U} = \mathbf{u}) \text{ by equation 1.3.4} \\
= \; & (\beta_2 + \beta_M m + \beta_{U_2} u_2 + \beta_{A_2} a_2) - (\beta_1 + \beta_M m + \beta_{U_1} u_1) \text{ by equations 1.3.1 and 1.3.2} \\
= \; & (\beta_2 - \beta_1) + \beta_{U_2} u_2 - \beta_{U_1} u_1 + \beta_{A_2} a_2
\end{aligned} \tag{1.3.5}$$

Equation 1.3.5 formalizes the intuition that, since M has the same effect on Y_1 and $Y_2(a_2)$, M has *no effect* on $\Delta Y_2(a_2)$. Thus the edge between M and $\Delta Y_2(a_2)$ in figure 1.3.2b is missing. We can also conclude from the absence of m in equation 1.3.5 that:

$$M \perp\!\!\!\perp \Delta Y_2(a_2) | \mathbf{U}, A_2. \tag{1.3.6}$$

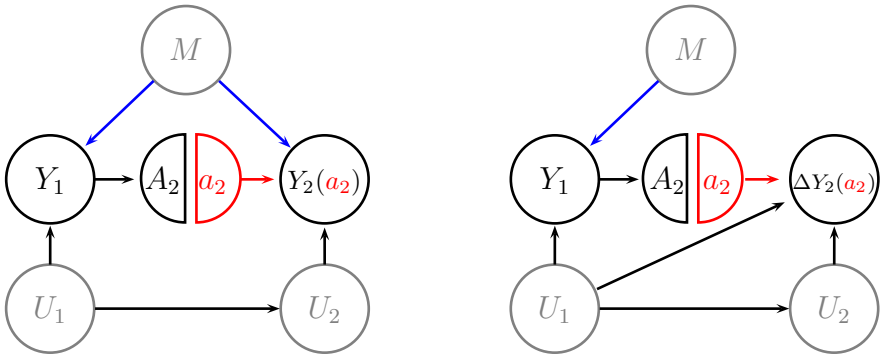
Finally, in appendix A, we establish that the assumption that the distribution P is Markov with respect to the SWIT in figure 1.3.2a implies that P is Markov with respect to the SWIT in figure 1.3.2b. This allow us to apply the rules of d-separation to figure 1.3.2b and establish that the causal relationships shown in this figure imply condition 1.2.1. There

are two paths from A_2 to $\Delta Y_2(a_2)$ in figure 1.3.2b: $A_2 \leftarrow M \rightarrow Y_1 \leftarrow U_1 \rightarrow \Delta Y_2(a_2)$ and $A_2 \leftarrow M \rightarrow Y_1 \leftarrow U_1 \rightarrow U_2 \rightarrow \Delta Y_2(a_2)$. Neither of these paths d-connect A_2 and $\Delta Y_2(a_2)$ given the empty set because Y_1 is a collider on each path. Therefore, A_2 and $\Delta Y_2(a_2)$ are d-separated given the empty set in figure 1.3.2b, so $A_2 \perp\!\!\!\perp \Delta Y_2(a_2) \; \forall \; a_2$ by theorem 1.2.1. In other words, condition 1.2.1 holds in general under this scenario, so the conclusion of statistician #1 is justified.

We can also apply the rules of d-separation to figure 1.3.2a and show that statistician #2 is generally incorrect in this scenario. Recall from condition 1.2.2 that statistician #2 is correct if $A_2 \perp\!\!\!\perp Y_2(a_2) \mid Y_1 \; \forall \; a_2$. But the path $A_2 \leftarrow M \rightarrow Y_2(a_2)$ in figure 1.3.2a d-connects A_2 and $Y_2(a_2)$ conditional on Y_1, so $A_2 \not\perp\!\!\!\perp Y_2(a_2) | Y_1 \; \forall \; a_2$ by theorem 1.2.2. This is an example of the oft-cited "omitted variable problem" in regression analysis (Wooldridge, 2012), and implies that condition 1.2.2 does not generally hold in this scenario and the estimate by statistician #2 is likely to be biased.

1.3.2 When is Statistician #2 Correct?

Suppose that the principal in our example chooses to assign students to the treatment group on the basis of observed test scores in year 1, Y_1. Since she wants to target the intervention to low-performing students, she and the other teachers in the school assign students to the treatment and control groups so that students with *lower* prior year test scores are *more likely* to be assigned to the treatment group (and students with higher prior year test scores are more likely to be assigned to the control group). Figure 1.3.3a shows the causal relationships between unobserved and observed variables in this scenario.



(a) SWIT for $Y_2(a_2)$ (b) SWIT for $\Delta Y_2(a_2)$

Figure 1.3.3: SWIT for statistician #2

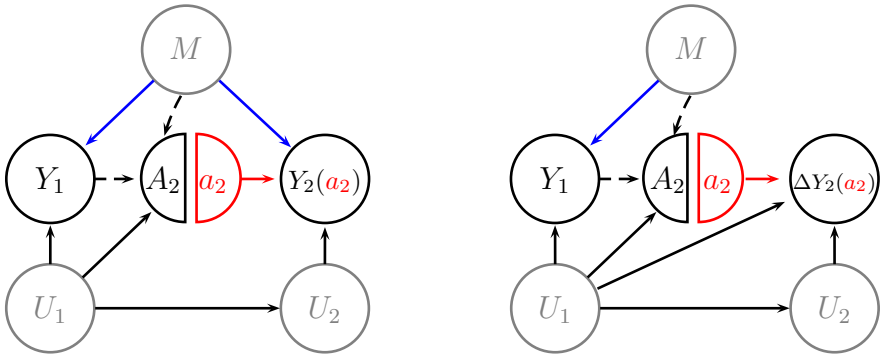
The edge between Y_1 and A_2 in figure 1.3.3a implies that students are selected for the intervention on the basis of prior year test scores. The absence of the edges between M and A_2 and between U_1 and A_2 implies that student's are *not* selected for the intervention on the basis of any unobserved (to the statisticians!) characteristic (unlike in the scenario described in section 1.3.1.)

We now apply the methods from section 1.2.3 to show that the relationships in figure 1.3.3a imply condition 1.2.2 ($A_2 \perp\!\!\!\perp Y_2(a_2) \mid Y_1 \; \forall \; a_2$), the condition that justifies the conclusion of statistician #2 (derived in section 1.2.1). There are three paths from A_2 to $Y_2(a_2)$ in figure 1.3.3: $A_2 \leftarrow Y_1 \leftarrow M \rightarrow Y_2(a_2)$; $A_2 \leftarrow Y_1 \leftarrow U_1 \rightarrow Y_2(a_2)$; and $A_2 \leftarrow Y_1 \leftarrow U_1 \rightarrow U_2 \rightarrow Y_2(a_2)$. Y_1 is a non-collider on each path, so none of these paths d-connect A_2 and $Y_2(a_2)$ given Y_1 in figure 1.3.3. Therefore, A_2 and $Y_2(a_2)$ are d-separated given Y_1 in figure 1.3.3, so $A_2 \perp\!\!\!\perp Y_2(a_2) \mid Y_1 \; \forall \; a_2$ by theorem 1.2.1. In other words, condition 1.2.2 holds in general under this scenario, so the conclusion of statistician #2 is justified.

We can also apply the rules of d-separation to figure 1.3.3 (derived from figure 1.3.3a as in section 1.3.1) and show that statistician #1 is generally incorrect in this scenario. Namely, the edge from Y_1 to A_2 in figure 1.3.3b creates a number of "backdoor paths" from A_2 to $\Delta Y_2(a_2)$ that d-connect A_2 and $\Delta Y_2(a_2)$ conditional on the empty set, meaning that $A_2 \not\perp\!\!\!\perp \Delta Y_2(a_2) \; \forall \; a_2$ by theorem 1.2.2. Thus condition 1.2.1 does not generally hold in this scenario and the estimate of statistician #1 is likely to be biased.

1.3.3 When are Both Statisticians Incorrect?

One conclusion that follows directly from our discussion in sections 1.3.1 and 1.3.2 is that both statisticians are generally incorrect if students are assigned to the intervention on the basis of *both* IQ score (scenario #1) and prior year test score (scenario #2). However, we consider a different (and perhaps more likely) scenario in this section. Suppose that the principal asks teachers to use their judgment to identify students who are struggling in class. Teachers use a variety of factors to identify students, including student motivation level, test scores that are unobserved to the statisticians, conversations with parents, and so on. The students who are judged to be struggling the most are assigned to the treatment group, while all other students are assigned to the control group. Figure 1.3.4a shows the causal relationships between unobserved and observed variables in this scenario.



(a) SWIT for $Y_2(a_2)$ (b) SWIT for $\Delta Y_2(a_2)$

Figure 1.3.4: SWITs describing scenario under which both statisticians are incorrect

The edge between U_1 and A_2 in figure 1.3.3a implies that students are selected for the intervention at least in part because of unobserved (to the statisticians!), time-variant student characteristics (motivation levels, unobserved test scores, parental support, etc.). We

draw the edges between M and A_2 and between Y_1 and A_2 as dashed edges to stress that in this scenario, it *doesn't matter* whether student ability or prior-year test scores are also used to assign students to the intervention: both statisticians are likely to be incorrect if students are selected at least in part because of unobserved, time-variant characteristics.

Specifically, the relationships in figure 1.3.4a imply that neither condition 1.2.1 ($A_2 \perp\!\!\!\perp \Delta Y_2(a_2) \; \forall \; a_2$) nor condition 1.2.2 ($A_2 \perp\!\!\!\perp Y_2(a_2) \mid Y_1 \; \forall \; a_2$) hold in this scenario. The path $A_2 \leftarrow U_1 \rightarrow Y_2(a_2)$ in figure 1.3.4a d-connects A_2 and $Y_2(a_2)$ conditional on Y_1 and thus implies that $A_2 \not\perp\!\!\!\perp Y_2(a_2)|Y_1 \; \forall \; a_2$ by theorem 1.2.2, so condition 1.2.2 does not hold and the estimate by statistician #2 is generally biased. Likewise, the path $A_2 \leftarrow U_1 \rightarrow U_2 \rightarrow \Delta Y_2(a_2)$ in figure 1.3.4b (derived from figure 1.3.4a as in section 1.3.1) d-connects A_2 and $\Delta Y_2(a_2)$ conditional on the empty set and thus implies that $A_2 \not\perp\!\!\!\perp \Delta Y_2(a_2) \; \forall \; a_2$ by theorem 1.2.2, so condition 1.2.1 does not hold and the estimate by statistician #1 is generally biased. This illustrates the well-known issue of selection on time-variant, unobserved variables, which is a major impediment to the evaluation of any non-randomized intervention and motivates more sophisticated estimation strategies such as instrumental variables (IV) estimation (e.g. Angrist et al., 1996).

1.4 Simulation study

We now describe an easily replicable simulation study that illustrates the three scenarios described in section 1.3. We have two goals in this simulation. The first is to reinforce our primary conclusions: statistician #1 is correct in the scenario described in section 1.3.1; statistician #2 is correct in the scenario described in section 1.3.2; and both statisticians are incorrect in the scenario described in section 1.3.3. To this end, we first describe a general simulation framework in section 1.4.1 that simulates from the SWIT for each scenario (figures 1.3.2a, 1.3.3a, and 1.3.4a). Interested readers can choose any parameter values they want within this general framework (including the true average treatment effect) and verify our primary conclusions.

Our second goal, however, is to demonstrate that each scenario in section 1.3 can produce data that are consistent with Lord's Paradox. With this in mind, for each scenario described in section 1.3, we specify parameter values in section 1.4.2 (including, most importantly, the true average treatment effect) that can be used within the general simulation framework from section 1.4.1 to generate data that are consistent with Lord's Paradox. We then describe the results of a simulation study that uses these specific parameter values.

1.4.1 General Simulation Framework

Figure 1.3.1 and equations 1.3.1 and 1.3.2 form the basis of all our simulations. For each student under each scenario, we simulate values for the time-constant effects M and year-1

time-variant effects U_1 as independent normal random variables with mean zero.

$$M \sim N(0, \sigma_M^2) \tag{1.4.1}$$
$$U_1 \sim N(0, \sigma_{U_1}^2) \tag{1.4.2}$$

From these two variables, we generate the value of the year-1 test score for each student as a normal random variable with the mean specified in equation 1.3.1.

$$Y_1|M = m, U_1 = u_1 \sim N(\beta_1 + \beta_M m + \beta_{U_1} u_1, \sigma_{Y_1}^2) \tag{1.4.3}$$

Values of the year-2 time-variant effects U_2 are simulated in two steps: we first simulate a change variable $\tilde{U}$ that describes how this variable changes from year 1 to year 2, and then simulate U_2 as a function of U_1, $\tilde{U}$, and a constant λ that describes the correlation between time-variant characteristics between the two years.

$$\tilde{U} \sim N(0, \sigma_{\tilde{U}}^2) \tag{1.4.4}$$
$$U_2|U_1 = u_1, \tilde{U} = \tilde{u} = \lambda u_1 + (1 - \lambda)\tilde{u} \tag{1.4.5}$$

We also simulate a treatment effect for each student, β_{A_2} in equation 1.3.2, as a normal variable centered at the average treatment effect β_{ATE}.

$$\beta_{A_2} \sim N(\beta_{ATE}, \sigma_{A_2}^2) \tag{1.4.6}$$

Finally, we generate the year-2 potential outcomes for each student as normal random variables with means specified in equation 1.3.2.

$$Y_2(0)|M = m, U_2 = u_2 \sim N(\beta_2 + \beta_M m + \beta_{U_2} u_2, \sigma_{Y_2}^2) \tag{1.4.7}$$
$$Y_2(1)|M = m, U_2 = u_2 \sim N(\beta_2 + \beta_M m + \beta_{U_2} u_2 + \beta_{A_2}, \sigma_{Y_2}^2) \tag{1.4.8}$$

The simulation framework above is completely independent of student assignment to the treatment ($A_2 = 1$) and control ($A_2 = 0$) groups. We now describe how we assign students to the treatment and control groups in our simulation for each scenario described in sections 1.3.1-1.3.3.

Treatment Assignment: Scenario #1

In the substantive scenario described in section 1.3.1, the principal assigns students in the lower half of the distribution of IQ scores to the treatment group. In the simulation for this scenario, we simply assign students with $M < 0$ to the treatment group:

$$(A_2|M = m, U_1 = u_1, Y_1 = y_1) = 1_{\{m<0\}} \tag{1.4.9}$$

Treatment Assignment: Scenario #2

In the substantive scenario described in section 1.3.2, the principal ensures that students with lower year 1 test scores are more likely to receive the intervention. In our simulation,

the probability that a student is placed in the treatment group is a monotonic decreasing function of Y_1.

$$(A_2|M = m, U_1 = u_1, Y_1 = y_1) \sim Bernoulli(\frac{exp(\alpha_{Y_1} y_1)}{exp(1 + \alpha_{Y_1} y_1)}),\ \alpha_{Y_1} < 0 \quad (1.4.10)$$

Treatment Assignment: Scenario #3

In the substantive scenario described in section 1.3.3, the principal assigns students to the intervention on the basis of time-variant characteristics in year 1. For simplicity, in the simulation for this scenario, we simply assign students with $U_1 < 0$ to the treatment group:

$$(A_2|M = m, U_1 = u_1, Y_1 = y_1) = 1_{\{u_1 < 0\}} \quad (1.4.11)$$

Regardless of how students are assigned to the treatment and control groups, the statisticians in our simulation observe one potential outcome for each student; $Y_2(0)$ if $A_2 = 0$, and $Y_2(1)$ if $A_2 = 1$. Using the observed test scores $Y_1 = y_1$ and $Y_2 = y_2$ from each simulation, we can produce an estimate for statistician #1 by subtracting the mean of $y_2 - y_1$ within the control group from the mean of $y_2 - y_1$ within the treatment group, and produce an estimate for statistician #2 by estimating linear regression model 1.1.1 and using the least squares estimate $\hat{\beta}_{A_2}$.

The general framework above contains 14 unspecified parameters in equations 1.4.1-1.4.8 and 1.4.10. Interested readers can select any parameter values and verify our general conclusions: statistician #1 produces an unbiased estimate in scenario #1 (section 1.3.1); statistician #2 produces an unbiased estimate in scenario #2 (section 1.3.2); and the estimates of both statisticians are biased in scenario #3 (section 1.3.3). In the next section, we go a step further and choose specific values of each of these parameters (including, most importantly, the true average treatment effect β_{ATE}) that generate data with the following (approximate) summary statistics that parallel properties 1.1.1-1.1.3 in section 1.1.2:

1. $-0.50 \approx \mathrm{E}(Y_1|A_2 = 1) \approx \mathrm{E}(Y_2|A_2 = 1) < \mathrm{E}(Y_1|A_2 = 0) \approx \mathrm{E}(Y_2|A_2 = 0) \approx 0.50$
2. $\mathrm{Var}(Y_1|A_2 = 1) \approx \mathrm{Var}(Y_2|A_2 = 1) \approx \mathrm{Var}(Y_1|A_2 = 0) \approx \mathrm{Var}(Y_2|A_2 = 0) \approx 0.87$
3. $\mathrm{Corr}(Y_1, Y_2|A_2 = 1) \approx \mathrm{Corr}(Y_1, Y_2|A_2 = 0) \approx 0.68$

In doing so, we demonstrate that the data in Lord's Paradox are consistent with three completely different explanations: (a) students were selected for the intervention on the basis of time-constant characteristics and there is **no** treatment effect (scenario #1); (b) students were selected for the intervention on the basis of prior year test scores and there is a **large negative** treatment effect (scenario #2); or (c) students were selected for the intervention on the basis of time-variant characteristics and there is a **modest negative** treatment effect (scenario #3).

1.4.2 Simulation Results

Our simulation study proceeds as follows. For each scenario, we perform 10,000 simulations with 1,000 students in each. For each simulation, we produce an estimate for statistician #1 and statistician #2, and then calculate the mean and standard deviation of these estimates over all 10,000 simulations for each scenario.

Table 1.4.1 contains the parameter values from equations 1.4.1-1.4.8 and 1.4.10 that we use in each scenario, the mean and standard deviation of the estimates produced by each statistician, and the coverage for each statistician's estimates (i.e., the percent of time the 95% confidence interval produced by each statistician includes the true average treatment effect we use in each scenario, shown in the first row). Unbiased estimates of the true average treatment effect are highlighted in green.

Scenario	**#1** (§1.3.1)	**#2** (§1.3.2)	**#3** (§1.3.3)
$\boldsymbol{\beta_{ATE}}$ (equation 1.4.6)	**0**	**-0.32**	**-0.23**
σ^2_M (equation 1.4.1)	0.4	0.4	0.4
$\sigma^2_{U_1}$ (equation 1.4.2)	0.4	0.4	0.4
β_M (equations 1.4.3, 1.4.7, and 1.4.8)	1	1	1
β_1 (equation 1.4.3)	0	0	0
β_{U_1} (equation 1.4.3)	1	1	1
$\sigma^2_{Y_1}$ (equation 1.4.3)	0.2	0.2	0.2
$\sigma^2_{\tilde{U}}$ (equation 1.4.4)	0.3	0.1	0.05
λ (equation 1.4.5)	0.825	0.5	0.6
$\sigma^2_{A_2}$ (equation 1.4.6)	0.1	0.1	0.2
β_2 (equations 1.4.7 and 1.4.8)	0	0.16	0.115
β_{U_2} (equations 1.4.7 and 1.4.8)	1	1	1
$\sigma^2_{Y_2}$ (equations 1.4.7 and 1.4.8)	0.2	0.2	0.2
α_{Y_1} (equation 1.4.10)	-	-1.35	-
Mean estimate, statistician #1	0.00	0.00	0.00
SD estimate, statistician #1	(0.04)	(0.04)	(0.04)
Coverage, statistician #1	94.82%	0.00%	0.06%
Mean estimate, statistician #2	-0.32	-0.32	-0.32
SD estimate, statistician #2	(0.05)	(0.05)	(0.05)
Coverage, statistician #2	0.00%	94.76%	55.07%

Table 1.4.1: Parameter values and estimates from simulation study.

Note that, since each simulation produces data with approximately the same summary statistics, the mean and standard deviations of the estimates produced by each statistician are the same across scenarios. We can see that statistician #1 produces an unbiased estimate in scenario #1, statistician #2 produces an unbiased estimate in scenario #2, and the estimates of both statisticians are biased in scenario #3, which verifies our primary conclusions.

But by relating this simulation study back to Lord's Paradox, we hope to illustrate a larger point. Namely, there is no way to know–without acquiring additional information– whether the data in our example come from a scenario in which students are assigned to the intervention on the basis of time-constant, time-invariant characteristics (scenario #1), prior year test scores (scenario #2), or time-variant characteristics (scenario #3). That is, not only is each statistician in our example correct under different scenarios, but there is no way to know *from the observed data* which statistician is correct (or whether both are incorrect). Thus our simulation illustrates a simple truth: it is not possible to infer causality from observational data without either: (a) making an assumption that is untestable from the observed data (e.g., assuming that treatment assignment is conditionally ignorable); or (b) acquiring more information about the treatment assignment process. The real-life example described in the next section illustrates this problem.

1.5 Preliminary Application to Special Education

Approximately 12% of all public school students in the United States receive special education services for a diagnosed disability, making special education the largest "targeted intervention" in American public schools. As we discuss in chapter 3, two large-scale, published papers have attempted to estimate the causal effect of special education services on student test performance, but interestingly, they come to very different conclusions. Hanushek et al. (2002) analyze longitudinal data from Texas public schools and report positive, statistically significant treatment effects of special education on student test performance; for example, students who transition into special education for a specific learning disability score 11% of a standard deviation higher in the years they receive special education services than in the years they do not. Morgan et al. (2010), on the other hand, analyze data from the Early Childhood Longitudinal Study and report negative or statistically insignificant treatment effects; for example, students in special education score 3.5 points lower in reading and 1.7 points lower in math, on average, than students in their matched non-special education sample.

These disparate findings could be due to different study settings, grade levels, or time periods, but the two sets of authors also use very different analytic strategies. Hanushek et al. (2002) estimate a student fixed effects model that estimates the average difference in student test score gains, all else equal, for students who transition into and out of special education between the years they were receiving special education services and the years they were not. This is analogous to the approach of statistician #1 in our example. Morgan et al. (2010), on the other hand, estimate a covariate adjustment model that estimates the average difference in performance, all else equal, between students in special education and "similar" students not in special education. This approach is analogous to statistician #2 in our example. We therefore argue that the disparate findings in Hanushek et al. (2002) and Morgan et al. (2010) could be an example of Lord's Paradox. In fact, Lord (1975) and Holland and Rubin (1983) specifically mention special education as one example of a targeted intervention that could lead to Lord's Paradox.

We illustrate this possibility with student-level data (discussed extensively in chapter 3) from Washington State public schools. For this preliminary application, we limit these data to the 2011-12 school year, and also limit the dataset to students who were *not* already receiving special education services at the beginning of the school year. Students who are placed into special education for a Specific Learning Disability (the largest disability category within special education) before the end of the school year are the "treatment group", while all other students serve as the control group.

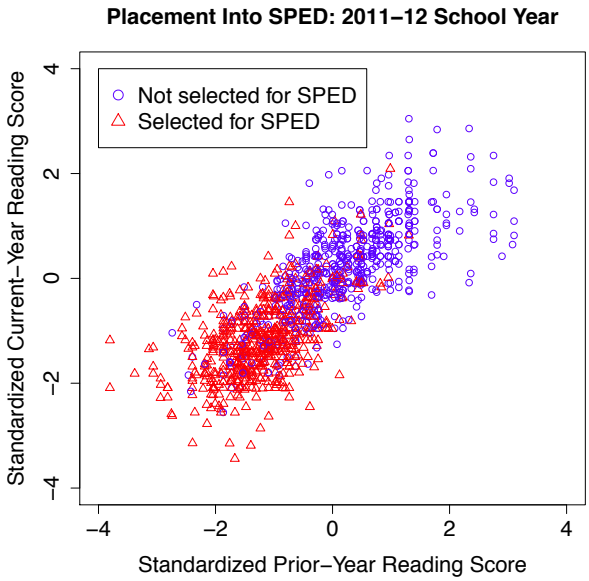


Figure 1.5.1: Observed data from special education placement in Washington State public schools

Figure 1.5.1 plots the prior-year and current-year test scores for the treatment and control groups (we plot scores for a random subset of the control group so the two groups have the same number of students in the figure). The summary statistics from these data are remarkably consistent with the motivating example in section 1.1.2: the average standardized scores within each group are approximately equal in years 1 and 2, and the average standardized scores in the treatment group are lower in both years than the average scores in the control group; the variances of the standardized scores are approximately equal for each combination of group and year; and the correlations between standardized scores in year 1 and year 2 for each group are approximately equal, and scores are not perfectly correlated between years. In fact, these real data are so consistent with the motivating example that we used them to illustrate Lord's Paradox in figures 1.1.1 and 1.1.2.

It is no surprise, then, that when we apply the methods of statisticians #1 and #2 to these real data, we see Lord's Paradox in action. That is, statistician #1 would conclude from these data that special education services have *no* impact on student test performance (see figure 1.1.2a), while statistician #2 would conclude that special education services have a large negative impact on student test performance (see figure 1.1.2b). These trends hold when we control for observable student characteristics, as in Hanushek et al. (2002) and Morgan et al. (2010): see chapter 3 for a full replication of the methods from these papers. Therefore, we are left with the same conundrum we described in section 1.4: there is no way to know (without gathering additional information about student assignment to special education) which set of authors is correct, or whether neither set of authors is correct.

That said, the conclusions from section 1.3–in which we used SWITs to describe the substantive scenarios under which each statistician in the motivating example is correct, and a substantive scenario under which neither statistician is correct–can be applied to this real-life example. Namely, if students are identified for special education on the basis of time-constant characteristics (such as an IQ score), then Hanushek et al. (2002) are likely to be correct (see figure 1.3.2) and Morgan et al. (2010) are likely to be incorrect (due to omitted variable bias). If students are identified for special education on the basis of the prior year test score (or other observed characteristics), then Morgan et al. (2010) are likely to be correct (see figure 1.3.3) and Hanushek et al. (2002) are likely to be incorrect. And if students are identified for special education on the basis of any unobserved characteristic that changes over time, then neither set of authors is likely to be correct (see figure 1.3.4). We provide a much more detailed investigation of the counterfactual conditions and substantive assumptions that justify each approach in chapter 3.

1.6 Conclusions

Researchers seeking to estimate the treatment effect of an intervention on an outcome variable often must use observational data in which the intervention is not randomly assigned to subjects. This chapter is intended as a cautionary tale of the risks inherent in drawing causal inferences from observational data, particularly when the intervention is targeted to certain types of subjects, potentially along unobserved dimensions. In this setting, the appropriate estimation method depends entirely on the assumptions that the researcher is willing to make, and as anyone who has ever attended a social science seminar is well aware, different researchers often have drastically different perspectives on which assumptions are appropriate for a given application.

This setting also makes it imperative for researchers to clearly communicate their assumptions to the reader or audience (not to mention to themselves)! A second goal of this paper is to illustrate the utility of Single World Intervention Templates (SWITs) as a means of communicating and verifying these assumptions. As we illustrate in sections 1.3.1-1.3.3, SWITs allow researchers not only to clarify the assumptions they have made, but connect these assumptions to the potential outcomes definition of confounding to illustrate how these assumptions lead to an unbiased (or biased!) estimate of the treatment effect. As we discuss

in section 1.3, though, it is nearly impossible to include all the potential confounders from a typical non-randomized educational intervention in a SWIT. This motivates our development of a new graphical framework, Single World Object Oriented Plates (SWOOPs), in chapter 2.

Chapter 2

Single World Object Oriented Plates (SWOOPs): A Graphical Framework for Causal Reasoning in Multivariate, Multilevel, and Longitudinal Settings

2.1 Introduction

In any discipline, the "gold standard" for estimating the causal effect of a treatment on an outcome is a large randomized experiment in which the treatment is randomized to subjects. This is true in education research as well, and a number of studies (e.g. Nye et al., 2000; Ritter et al., 2007) have used randomization to produce credible estimates of the causal effects of various educational interventions on student outcomes. However, randomization is often not feasible in school settings. For example, it may not be possible (or ethical!) for a school to randomly assign federally-mandated interventions like special education to students, for a district to deprive resources from a random subset of schools, or for a state to randomly assign districts to one of two funding systems. Therefore, education researchers interested in the causal effects of these treatments on outcomes like student achievement typically estimate these effects from observational data in which the treatment has *not* been randomized to students.

In chapter 1, we illustrate the utility of graphical methods for communicating and verifying the assumptions that justify causal conclusions from observational data. The problem with the simplified example in that chapter, though, is that a typical non-randomized educational intervention has a much larger set of potential confounders; that is, variables that influence student educational outcomes but could also influence the probability that students receive the intervention. Figure 2.1.1 shows a Directed Acyclic Graph (DAG) that includes some confounders ($C_1, \ldots, C_8$) that could complicate estimating the causal effect of some non-randomized intervention A on an outcome Y. The black edge from each C_i to the outcome Y indicates that each of these variables is assumed to influence the outcome, while the blue arrow from each variable to the intervention A indicates that each variable *could*

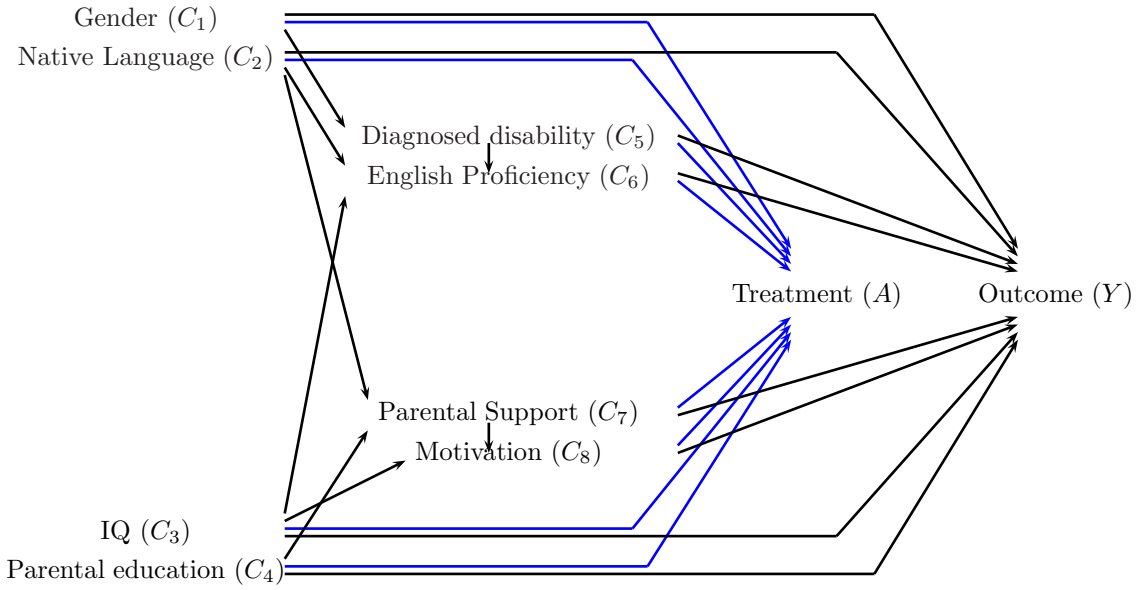


Figure 2.1.1: Conceptual figure of non-randomized educational intervention

also potentially influence student assignment to the intervention. Black edges between the C_i variables represent potential causal relationships between these variables.

If the DAG in figure 2.1.1 represented all relationships between all variables (observed and unobserved) in this hypothetical study, we could follow the same procedure outlined in the previous chapter: represent this DAG as a Single World Intervention Template (SWIT, Richardson and Robins, 2013), apply the rules of d-separation (Pearl, 2000) to the SWIT, and establish the counterfactual independence conditions that validate causal conclusions in this study. Unfortunately, the DAG in 2.1.1 is insufficient because it only shows a small subset of potential confounders in a typical non-randomized educational intervention, and for only one student, in one school, and in one year. Given the number of students, schools, years, and potential confounders in a typical non-randomized educational intervention, while it is possible in theory to represent all the necessary relationships in a DAG, the resulting graph would likely not be useful in practice.

In this paper, we introduce a new graphical framework–Single World Object Oriented Plates, or SWOOPs–that addresses these limitations. We develop SWOOPs in section 2.2, and prove that conditional independence relationships in SWOOPs imply conditional independence relationships in any underlying DAG. The utility of SWOOPs, outlined in section 2.3, is that they permit causal reasoning in the multivariate, multilevel, and longitudinal settings that are common in education research. As an application, we consider value-added models (VAMs) of teacher effectiveness in section 2.4, and show that SWOOPs can be used to connect the substantive conditions (Rothstein, 2009) and the counterfactual conditions

(Reardon and Raudenbush, 2009) that justify causal conclusions from different VAMs. We offer some concluding thoughts and potential extensions in section 2.5.

2.2 Development

We develop SWOOPs from DAGs in three steps, each of which borrows concepts from a different strand of graphical theory. First, we build on the literature on object-oriented networks (e.g. Koller and Pfeffer, 1997) and develop Object-Oriented Graphs (OOGs) in section 2.2.1. We then introduce "plates" (Buntine, 1994; Heckerman et al., 2007) in section 2.2.2, and show how plates can dramatically simplify relationships in OOGs with repeated structures. Finally, in section 2.2.3 we combine these ideas with recent developments from Richardson and Robins (2013), who include potential outcomes on a graph via a "node-splitting" operation on intervention nodes, to define Single-World Object-Oriented Plates (SWOOPs).

2.2.1 Objects

Definitions

Although our development of object-oriented graphs is more similar to Bangsø and Wuillemin (2000) and Dawid et al. (2007), we use the terminology of Koller and Pfeffer (1997) throughout this section because Heckerman et al. (2007) use the same terminology to define plates (discussed in section 2.2.2). Following Heckerman et al. (2007), we distinguish between *entities* and *entity classes*.

Definition 2.2.1 *An **entity** is a unit of analysis in a specific dataset.*

Definition 2.2.2 *An **entity class** is a set of entities without specification of the entities in the set.*

The conceptual diagram in figure 2.1.1 describes relationships for the entity class "student", rather than relationships for any specific entity (e.g., for an individual student in a dataset). Each of the nodes in figure 2.1.1 are called *attributes*.

Definition 2.2.3 *An **attribute** is a variable describing some property of an entity, represented by a single node in a DAG.*

The key development in this section is to use a mapping $\mathcal{M}_n$ to arrange attributes $X_1, \ldots, X_N$ into different *objects* $O_1, \ldots, O_n,\ n \leq N$.

Definition 2.2.4 *An **object** is a collection of one or more attributes from a DAG that is represented as a single node in a OOG. The collection of attributes in a DAG mapped to an object O_i by a mapping $\mathcal{M}_n$ is denoted $\mathbf{X}_{O_i}$. If $\mathbf{X}_{O_i}$ contains only one attribute, then O_i is a **simple object**. Otherwise, O_i is a **complex object**.*

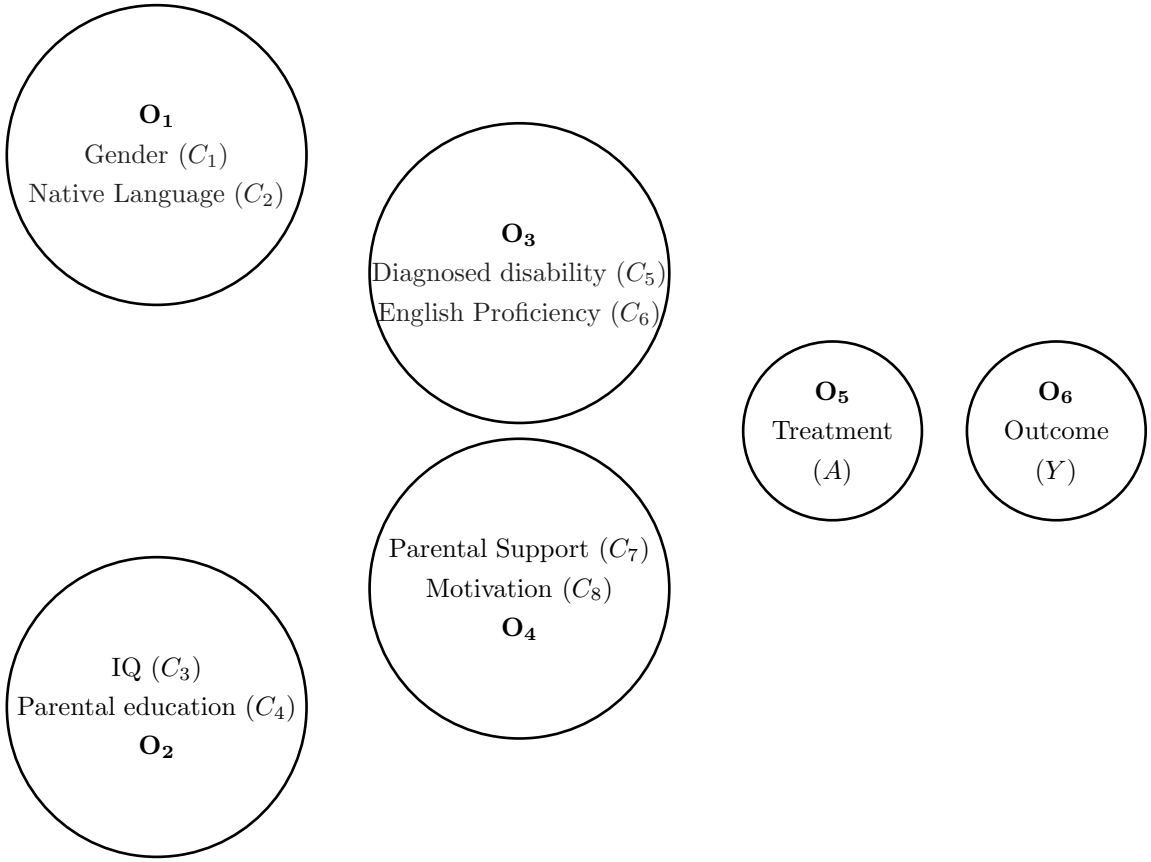


Figure 2.2.1: Example of mapping $\mathcal{M}_6$

As one example, we can define a mapping $\mathcal{M}_6$ that maps the $N = 10$ student attributes in figure 2.1.1 into $n = 6$ objects, shown in figure 2.2.1. This particular mapping will be central to our subsequent discussion: $\mathbf{O_1}$ is the collection of attributes that are *time-invariant* and *observed* to researchers; $\mathbf{O_2}$ is the collection of attributes that are *time-invariant* and *unobserved* to researchers; $\mathbf{O_3}$ is the collection of attributes that are *time-variant* and *observed* to researchers; $\mathbf{O_4}$ is the collection of attributes that are *time-variant* and *unobserved* to researchers; $\mathbf{O_5}$ contains the treatment variable; and $\mathbf{O_6}$ contains the outcome variable. O_5 and O_6 are both simple objects (because they contain only one attribute), while O_1, O_2, O_3, and O_4 are all complex objects (because they contain more than one attribute).

Given a DAG $\mathcal{G}$ with N attributes, a mapping $\mathcal{M}_n$ of the N attributes in $\mathcal{G}$ into $n \leq N$ objects must satisfy property 2.2.1 for the resulting graph to be an *object-oriented graph (OOG)*.

Property 2.2.1 *A mapping $\mathcal{M}_n$ of the N attributes in a DAG $\mathcal{G}$ into $n \leq N$ objects $O_1, \ldots, O_n$ results in an **object-oriented graph (OOG)** $\mathcal{G}^{\mathcal{M}_n}$ if and only if, every time attributes X_i and X_j with an edge $X_i \to X_j$ in $\mathcal{G}$ are mapped into different objects O_i and O_j (so $X_i \in \mathbf{X}_{O_i}$ and $X_j \in \mathbf{X}_{O_j}$), there is no directed edge from any attribute in $\mathbf{X}_{O_j}$ to any attribute in $\mathbf{X}_{O_i}$ in $\mathcal{G}$.*

Property 2.2.1 ensures that the resulting OOG $\mathcal{G}^{\mathcal{M}_n}$ is itself a DAG (with nodes $O_1, \ldots, O_n$). Note that the mapping $\mathcal{M}_6$ shown in figure 2.2.1 satisfies property 2.2.1.

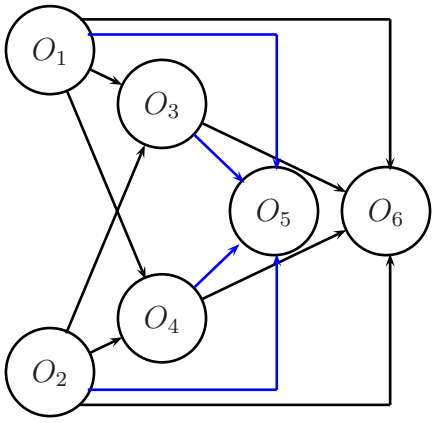


Figure 2.2.2: Example of Object-Oriented Graph (OOG)

If the mapping $\mathcal{M}_n$ satisfies property 2.2.1, we can derive the edges in the resulting OOG $\mathcal{G}^{\mathcal{M}_n}$ by considering each pair of objects $\{O_i, O_j\}$. Definition 2.2.5 allows us to determine the edge from O_i to O_j in $\mathcal{G}^{\mathcal{M}_n}$ from the edges in $\mathcal{G}$.

Definition 2.2.5 *An edge from object O_i to O_j in an OOG $\mathcal{G}^{\mathcal{M}_n}$ is **directed from object** O_i **to object** O_j ($O_i \to O_j$) if there is at least one directed edge from an attribute in $\mathbf{X}_{\mathbf{O_i}}$ to an attribute in $\mathbf{X}_{\mathbf{O_j}}$ in $\mathcal{G}$.*

Definition 2.2.5 implies that an edge is missing between objects O_i and O_j if and only if there is no directed edge from any attribute in $\mathbf{X}_{\mathbf{O_i}}$ to any attribute in $\mathbf{X}_{\mathbf{O_j}}$ in $\mathcal{G}$, and no directed edge from any attribute in $\mathbf{X}_{\mathbf{O_j}}$ to any attribute in $\mathbf{X}_{\mathbf{O_i}}$ in $\mathcal{G}$. Figure 2.2.2 shows the resulting edges in the OOG $\mathcal{G}^{\mathcal{M}_6}$ (from applying definition 2.2.5 to the DAG $\mathcal{G}$ in figure 2.1.1 and mapping $\mathcal{M}_6$ in figure 2.2.1).

An alternative procedure for deriving the edges in an OOG is to derive paths in an OOG from the paths in a DAG.

Definition 2.2.6 *Given a path π in a DAG $\mathcal{G}$ consisting of vertices $V_1, \ldots, V_k$, the **derived path** $\pi^{\mathcal{M}_n}$ in $\mathcal{G}^{\mathcal{M}_n}$ can be derived recursively as follows:*

1. *The first vertex in $\pi^{\mathcal{M}_n}$ is the object O_i to which $\mathcal{M}_n$ maps V_1.*
2. *For each subsequent vertex in π, if V_j is mapped by $\mathcal{M}_n$ to the same object as V_{j-1}, proceed to the next vertex. If V_j is mapped by $\mathcal{M}_n$ to a different object O_j than V_{j-1}, then O_j becomes the next vertex in $\pi^{\mathcal{M}_n}$, and is connected to the previous vertex in $\pi^{\mathcal{M}_n}$ by the edge between V_{j-1} and V_j.*

As an example, consider the path $\pi : C_4 \to C_7 \to C_8 \to Y$ in the DAG $\mathcal{G}$ in figure 2.1.1. C_4 is mapped to O_2 by $\mathcal{M}_6$, so the first vertex in $\pi^{\mathcal{M}_6}$ is O_2. C_7 is mapped to O_4 by $\mathcal{M}_6$, so the next vertex in $\pi^{\mathcal{M}_6}$ is O_4, and is connected to O_2 by $\to$ (because $C_4 \to C_7$ in π). On the other hand, C_8 is also mapped to O_4, so we can proceed to the next vertex. Finally, Y is mapped to O_6 by $\mathcal{M}_6$, so the next vertex in $\pi^{\mathcal{M}_6}$ is O_6, and is connected to O_4 by $\to$ (because $C_8 \to Y$ in π). This completes the derivation of the derived path; that is, $\pi^{\mathcal{M}_6}$ is

$O_2 \to O_4 \to O_6$.

Multiple paths in $\mathcal{G}$ can result in the same derived path in $\mathcal{G}^{\mathcal{M}_n}$ (e.g., the path $C_3 \to C_8 \to Y$ results in the same derived path as the example above). Further, not every path in an OOG $\mathcal{G}^{\mathcal{M}_n}$ can be derived from a path in $\mathcal{G}$ (e.g., suppose the edge between C_7 and C_8 and the edge between C_7 and Y were missing in figure 2.1.1. Then $\mathcal{G}^{\mathcal{M}_n}$ would still have the path $O_2 \to O_4 \to Y$, but this path could not be derived from any path in $\mathcal{G}$).

We now prove two lemmas about derived paths before proceeding to a discussion of d-separation in OOGs.

Lemma 2.2.1 *If a path π^* in an OOG $\mathcal{G}^{\mathcal{M}_n}$ contains a non-collider O_{NC}, then either there is no path π in $\mathcal{G}$ that results in the derived path π^*, or each path in $\mathcal{G}$ that results in the derived path π^* contains a non-collider $X_{NC} \in \mathbf{X_{O_{NC}}}$.*

Pf. (by contradiction) Assume there exists a path π in $\mathcal{G}$ such that $\pi^{\mathcal{M}_n} = \pi^*$, and π does not contain a non-collider $X_{NC} \in \mathbf{X_{O_{NC}}}$. $\pi^{\mathcal{M}_n}$ has of one of three forms: (1) $\ldots O_i \to O_{NC} \to O_j \ldots$; (2) $\ldots O_i \leftarrow O_{NC} \to O_j \ldots$; or (3) $\ldots O_i \leftarrow O_{NC} \leftarrow O_j \ldots$. The proof in each case is nearly identical, so WLOG, consider just the first case. If $\pi^{\mathcal{M}_n}$ is of the form $\ldots O_i \to O_{NC} \to O_j \ldots$, then π is of the form $\ldots X_i \to X_1^* \ldots X_2^* \to X_j \ldots$, where all vertices between X_1^* and X_2^* (including these vertices) are in $\mathbf{X_{O_{NC}}}$ (note that, if there was only one such vertex, it would trivially be a non-collider with respect to π). X_1^* cannot be a non-collider with respect to the path π, so π is of the form $\ldots X_i \to X_1^* \leftarrow \ldots X_2^* \leftarrow X_j \ldots$. But then the next vertex in the path (which is also in $\mathbf{X_{O_{NC}}}$) must be a non-collider with respect to π. $\to\leftarrow$. $\square$

Lemma 2.2.2 *If a path π^* in an OOG $\mathcal{G}^{\mathcal{M}_n}$ contains a collider O_C, then either there is no path π in $\mathcal{G}$ that results in the derived path π^*, or each path in $\mathcal{G}$ that results in the derived path π^* contains a collider $X_C \in \mathbf{X_{O_C}}$.*

Pf. (by contradiction) Assume there exists a path π in $\mathcal{G}$ such that $\pi^{\mathcal{M}_n} = \pi^*$, and π does not contain a collider $X_C \in \mathbf{X_{O_C}}$. $\pi^{\mathcal{M}_n}$ is of the form $\ldots O_i \to O_C \leftarrow O_j \ldots$, so π is of the form $\ldots X_i \to X_1^* \ldots X_2^* \leftarrow X_j \ldots$, where all vertices between X_1^* and X_2^* (including these vertices) are in $\mathbf{X_{O_C}}$. X_1^* cannot be a collider with respect to the path π, so π is of the form $\ldots X_i \to X_1^* \to \ldots X_2^* \leftarrow X_j \ldots$. But this is true of each of the attributes between X_1^* and X_2^* in π, so π is of the form $\ldots X_i \to X_1^* \to \ldots \to X_2^* \leftarrow X_j \ldots$. But then X_2^* is a collider with respect to π. $\to\leftarrow$. $\square$

D-separation in OOGs

As we describe above, each DAG $\mathcal{G}$ and mapping $\mathcal{M}_n$ that satisfies property 2.2.1 implies a unique OOG $\mathcal{G}^{\mathcal{M}_n}$. However, if we are given a mapping $\mathcal{M}_n$ and OOG $\mathcal{G}^{\mathcal{M}_n}$, there are often a number of different "underlying" DAGs $\mathcal{G}$. This is easy to see from the OOG in figure 2.2.2; this OOG tells us nothing about the relationships between the attributes *within* $\mathbf{X_{O_1}}$, $\mathbf{X_{O_2}}$, etc. in $\mathcal{G}$, and nothing about the *number* of edges between attributes in $\mathbf{X_{O_1}}$, $\mathbf{X_{O_2}}$, and the intervention and outcome nodes in $\mathcal{G}$.

However, the purpose of this section is to demonstrate that we can still draw conclusions about the conditional independence between sets of attributes in *any* underlying DAG $\mathcal{G}$ given d-separation (definition 1.2.8 in chapter 1) between the objects that contain these attributes in an OOG $\mathcal{G}^{\mathcal{M}_n}$. We begin by proving two lemmas relating an OOG $\mathcal{G}^{\mathcal{M}_n}$ to any underlying DAG $\mathcal{G}$.

Lemma 2.2.3 *If no path connects objects O_i and O_j in an OOG $\mathcal{G}^{\mathcal{M}_n}$, then no path connects any attribute in $\mathbf{X_{O_i}}$ to any attribute in $\mathbf{X_{O_j}}$ in any underlying DAG $\mathcal{G}$.*

Pf. (by contradiction) Suppose that there exists an underlying DAG $\mathcal{G}$ with a path π connecting an attribute $X_i \in O_i$ and an attribute $X_j \in O_j$. Consider the derived path $\pi^{\mathcal{M}_n}$ in $\mathcal{G}^{\mathcal{M}_n}$. By definitions and 2.2.5 and 2.2.6, $\pi^{\mathcal{M}_n}$ must be of the form $O_i \ldots O_j$, and thus connects O_i and O_j in $\mathcal{G}^{\mathcal{M}_n}$. $\rightarrow\leftarrow$. $\square$

Lemma 2.2.4 *Let $\mathcal{G}^{\mathcal{M}_n}$ be an OOG with objects $O_1, \ldots, O_n$. Let O_i and O_j be any pair of distinct objects, and let S be any subset of the other objects (including the empty set). Suppose that every path between O_i and O_j in $\mathcal{G}^{\mathcal{M}_n}$ either: (a) has a non-collider in S; or (b) has a collider that is not an ancestor of S (and not in S). Then every path π between an attribute in $\mathbf{X_{O_i}}$ and an attribute in $\mathbf{X_{O_j}}$ in any underlying DAG $\mathcal{G}$ satisfies one of two properties: (c) π has a non-collider in $\mathbf{X_S}$; or (d) π has a collider that is not an ancestor of $\mathbf{X_S}$ (and not in $\mathbf{X_S}$).*

Pf. (by contradiction) Suppose there is a path π between an attribute $X_i \in \mathbf{X_{O_i}}$ and an attribute $X_j \in \mathbf{X_{O_j}}$ in *any* underlying DAG $\mathcal{G}$ that does not satisfy property (c) or property (d). Then there are two cases:

Case 1: Suppose $\pi^{\mathcal{M}_n}$ satisfies property (a); i.e., $\pi^{\mathcal{M}_n}$ has a non-collider $O_{NC} \in S$. By lemma 2.2.1, π must contain a non-collider $X_{NC} \in \mathbf{X_{O_{NC}}}$, which implies that π has a non-collider in $\mathbf{X_S}$. Thus π must satisfy property (c).

Case 2: Suppose $\pi^{\mathcal{M}_n}$ satisfies property (b); i.e., $\pi^{\mathcal{M}_n}$ has a collider O_C that is not an ancestor of S and not in S. By lemma 2.2.2, π must have a collider $X_C \in \mathbf{X_{O^*}}$. This collider cannot be an ancestor of $\mathbf{X_S}$ or be in $\mathbf{X_S}$, because if it were, then O_C would be an ancestor of S or be in S. Thus π must satisfy property (d).

Since $\pi^{\mathcal{M}_n}$ must satisfy either property (a) or property (b), π must satisfy property (c) or property (d). $\rightarrow\leftarrow$. $\square$

We now use lemmas 2.2.3 and 2.2.4 to prove theorem 2.2.1.

Theorem 2.2.1 *Let $\mathcal{G}^{\mathcal{M}_n}$ be an OOG with objects $O_1, \ldots, O_n$. Let O_i and O_j be any pair of distinct objects, and let S be any subset of the other objects (including the empty set). If O_i and O_j are d-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in any underlying DAG $\mathcal{G}$.*

Pf. If O_i and O_j are d-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then no path in $\mathcal{G}^{\mathcal{M}_n}$ d-connects O_i and O_j given S. Then there are two possibilities:

Case 1: If no path in $\mathcal{G}^{\mathcal{M}_n}$ connects O_i and O_j, then by lemma 2.2.3, no path connects any attribute in $\mathbf{X_{O_i}}$ and any attribute in $\mathbf{X_{O_j}}$ in any underlying DAG $\mathcal{G}$. Thus each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in any underlying DAG $\mathcal{G}$.

Case 2: If there is at least one path in $\mathcal{G}^{\mathcal{M}_n}$ connecting O_i and O_j, then each of these paths: (a) has a non-collider in S; or (b) has a collider that is not an ancestor of S (and not in S). Consider any underlying DAG $\mathcal{G}$, and consider the set of all paths between attributes in $\mathbf{X_{O_i}}$ and attributes in $\mathbf{X_{O_j}}$ in $\mathcal{G}$ (if any exist). By lemma 2.2.4, each of these paths either has a non-collider in $\mathbf{X_S}$ or has a collider that is not an ancestor of $\mathbf{X_S}$ (and not in $\mathbf{X_S}$). Thus no path d-connects any attribute in $\mathbf{X_{O_i}}$ and any attribute in $\mathbf{X_{O_i}}$ given $\mathbf{X_S}$ in $\mathcal{G}$, so each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in $\mathcal{G}$. This is true of any underlying DAG $\mathcal{G}$, so each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in any underlying DAG $\mathcal{G}$.

Thus in each case, each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in any underlying DAG $\mathcal{G}$. □

Theorem 2.2.1 is extremely general; importantly, *no knowledge* of the structure of the underlying DAG $\mathcal{G}$ is necessary to infer d-separation between sets of attributes in $\mathcal{G}$ given d-separation between the object that contain these attributes in $\mathcal{G}^{\mathcal{M}_n}$. This generality allows us to combine theorem 2.2.1 with theorem 1.2.1 from chapter 1 to relate d-separation in an OOG $\mathcal{G}^{\mathcal{M}_n}$ to conditional independence between sets of variable in any underlying DAG $\mathcal{G}$.

Theorem 2.2.2 *Given an underlying DAG $\mathcal{G}$ and an OOG $\mathcal{G}^{\mathcal{M}_n}$ with objects $O_1, \ldots, O_n$, let O_i and O_j be any pair of distinct objects in $\mathcal{G}^{\mathcal{M}_n}$, and let S be any subset of the other objects in $\mathcal{G}^{\mathcal{M}_n}$ (including the empty set). In any distribution P that is Markov with respect to $\mathcal{G}$, if objects O_i and O_j are d-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then $\mathbf{X_{O_i}} \perp\!\!\!\perp \mathbf{X_{O_j}} | \mathbf{X_S}$ in $\mathcal{G}$ and P.*

Pf. By theorem 2.2.1, if objects O_i and O_j are d-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then each attribute in $\mathbf{X_{O_i}}$ is d-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in $\mathcal{G}$. By theorem 1.2.1 from chapter 1, then, $\mathbf{X_{O_i}} \perp\!\!\!\perp \mathbf{X_{O_j}} | \mathbf{X_S}$ in $\mathcal{G}$ and P. □

Theorem 2.2.2 is the central result of this chapter, as it will allow us to draw conclusions about relationship between attributes in a DAG from the simplified relationships shown in an OOG.

Deterministic objects

In our subsequent discussion, we will at times consider DAGs that include attributes that are deterministic functions of other attributes in the DAG (e.g., school-level characteristics are a deterministic function of an indicator for each individual school). Geiger et al. (1990) and Spirtes et al. (2000) demonstrate that the rules of d-separation are not sufficient to

establish conditional independence relationships in a DAG with these types of "deterministic attributes."

Definition 2.2.7 *An attribute X in a DAG $\mathcal{G}$ is a* ***deterministic attribute*** *(and represented by a node with a double circle) if it is a deterministic function of its parents in $\mathcal{G}$.*

Definition 2.2.8 *An attribute X_i in a DAG $\mathcal{G}$ is* ***functionally determined*** *by a set of attributes $\mathbf{X}$ if and only if $X_i \in \mathbf{X}$ or X_i is a deterministic attribute and all its parents are functionally determined by $\mathbf{X}$.*

Geiger et al. (1990) and Spirtes et al. (2000) use definitions 2.2.7 and 2.2.8 to define D-separation (note the capital "D" distinguishes this rule from the typical rule of (lower-case) d-separation in definition 1.2.8).

Definition 2.2.9 *In a DAG $\mathcal{G}$ that contains deterministic attributes, a path π* ***D-connects*** *vertices A and B conditional on a set S in $\mathcal{G}$ if A and B are the endpoints of π, every non-collider on π is not functionally determined by S, and every collider on π is an ancestor of S (or is in S). If there is no path D-connecting A and B given S in $\mathcal{G}$, then A and B are* ***D-separated*** *given S in $\mathcal{G}$.*

Now, suppose $\mathcal{G}$ is a DAG that includes deterministic attributes. In this scenario, we have to place additional restrictions on the mapping $\mathcal{M}_n$ to ensure that definition 2.2.8 is "portable" across DAGs and OOGs. Specifically, we will require that $\mathcal{M}_n$ map deterministic attributes only to "deterministic objects" that contain deterministic attributes with the same parents, which in turn must be mapped to their own objects.

Property 2.2.2 *Suppose a DAG $\mathcal{G}$ has deterministic attributes $Z_1, \ldots, Z_{N_Z}$, each of which has parents $I_1, \ldots, I_{N_I}$. In addition to property 2.2.1, a mapping $\mathcal{M}_n$ of the N attributes in $\mathcal{G}$ into $n \leq N$ objects $O_1, \ldots, O_n$ results in an OOG $\mathcal{G}^{\mathcal{M}_n}$ if and only if, when $\mathcal{M}_n$ maps $Z_i \in \{Z_1, \ldots, Z_{N_Z}\}$ to an object O_i, all attributes in $\mathbf{X}_{\mathbf{O_i}}$ are also deterministic attributes with the same parents as Z_i in $\mathcal{G}$, and the parents of Z_i are mapped to their own object $O_I \neq O_i$. The object O_i is called a* ***deterministic object****, and is denoted by a node with a double circle in $\mathcal{G}^{\mathcal{M}_n}$.*

As a simple example, suppose that the treatment in figure 2.1.1 is assigned to a student if and only if the student has a given set of time-invariant, observable characteristics (e.g., a given gender, native language, etc.). Then O_5 from the OOG in figure 2.2.2 becomes a deterministic object with parent O_1. The resulting OOG is shown in figure 2.2.3 (note that property 2.2.2 holds in this example).

To establish conditional independencies from an OOG like the one in figure 2.2.3, we need to prove a version of theorem 2.2.1 for the case where $\mathcal{G}^{\mathcal{M}_n}$ contains deterministic objects.

Theorem 2.2.3 *Let $\mathcal{G}^{\mathcal{M}_n}$ be an OOG with objects $O_1, \ldots, O_n$, some of which are deterministic objects. Let O_i and O_j be any pair of distinct non-deterministic objects, and let S be any subset of the other objects (including the empty set). If O_i and O_j are D-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then each attribute in $\mathbf{X}_{\mathbf{O_i}}$ is D-separated from each attribute in $\mathbf{X}_{\mathbf{O_j}}$ given $\mathbf{X}_{\mathbf{S}}$ in any underlying DAG $\mathcal{G}$.*

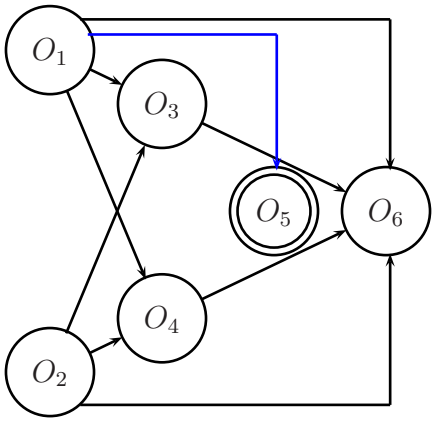


Figure 2.2.3: Example of OOG with deterministic object

Instead of going through the whole proof, we simply note that the only case not covered by the proof of theorem 2.2.1 is when the set S in $\mathcal{G}^{\mathcal{M}_n}$ functionally determines some object O_k. However, if we define $det(S)$ as the set of objects in $\mathcal{G}^{\mathcal{M}_n}$ functionally determined by S, then property 2.2.2 ensures that that $det(\mathbf{X_S}) = \mathbf{X_{det(S)}}$ in $\mathcal{G}$. This allows us to prove a modified version of lemma 2.2.4 that is aligned with definition 2.2.9 (this in turn is sufficient to prove theorem 2.2.3).

Lemma 2.2.5 *Let $\mathcal{G}^{\mathcal{M}_n}$ be an OOG with objects $O_1, \ldots, O_n$, some of which are deterministic objects. Let O_i and O_j be any pair of distinct non-deterministic objects, and let S be any subset of the other objects (including the empty set). Suppose that every path between O_i and O_j in $\mathcal{G}^{\mathcal{M}_n}$ either: (a) has a non-collider in $det(S)$; or (b) has a collider that is not an ancestor of S (and not in S). Then every path π between an attribute in $\mathbf{X_{O_i}}$ and an attribute in $\mathbf{X_{O_j}}$ in any underlying DAG $\mathcal{G}$ satisfies one of two properties: (c) π has a non-collider in $\mathbf{X_{det(S)}} = det(\mathbf{X_S})$; or (d) π has a collider that is not an ancestor of $\mathbf{X_S}$ (and not in $\mathbf{X_S}$).*

Pf. (by contradiction) Suppose there is a path π between an attribute $X_i \in \mathbf{X_{O_i}}$ and an attribute $X_j \in \mathbf{X_{O_j}}$ in *any* underlying DAG $\mathcal{G}$ that does not satisfy property (c) or property (d). The proof of lemma 2.2.4 shows that property (b) implies property (d), so we only need to consider property (a). Suppose $\pi^{\mathcal{M}_n}$ satisfies property (a); i.e., $\pi^{\mathcal{M}_n}$ has a non-collider $O_{NC} \in det(S)$. By lemma 2.2.1, π must contain a non-collider $X_{NC} \in \mathbf{X_{O_{NC}}}$, which implies that π has a non-collider in $\mathbf{X_{det(S)}} = det(\mathbf{X_S})$. Thus π must satisfy property (c).

Since $\pi^{\mathcal{M}_n}$ must satisfy either property (a) or property (b), π must satisfy property (c) or property (d). $\rightarrow\leftarrow$. $\square$

Theorem 2.2.3 allows us to use D-separation in an OOG with deterministic objects to infer D-separation between attributes in any underlying DAG with deterministic attributes. We can combine theorem 2.2.3 with the following theorem from Geiger (1990):

Theorem 2.2.4 *In any distribution P that is Markov with respect to $\mathcal{G}$, if A and B are D-separated given $\mathbf{C}$ in $\mathcal{G}$ then $A \perp\!\!\!\perp B|\mathbf{C}$ in P.*

This allows us to prove a version of theorem 2.2.2 for OOGs with deterministic objects:

Theorem 2.2.5 *Given an underlying DAG $\mathcal{G}$ and an OOG $\mathcal{G}^{\mathcal{M}_n}$ with objects $O_1, \ldots, O_n$, some of which are deterministic objects, let O_i and O_j be any pair of distinct non-deterministic objects in $\mathcal{G}^{\mathcal{M}_n}$, and let S be any subset of the other objects in $\mathcal{G}^{\mathcal{M}_n}$ (including the empty set). In any distribution P that is Markov with respect to $\mathcal{G}$, if objects O_i and O_j are D-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then $\mathbf{X_{O_i}} \perp\!\!\!\perp \mathbf{X_{O_j}} | \mathbf{X_S}$ in $\mathcal{G}$ and P.*

Pf. By theorem 2.2.3, if objects O_i and O_j are D-separated given S in $\mathcal{G}^{\mathcal{M}_n}$, then each attribute in $\mathbf{X_{O_i}}$ is D-separated from each attribute in $\mathbf{X_{O_j}}$ given $\mathbf{X_S}$ in $\mathcal{G}$. By theorem 2.2.4, then, $\mathbf{X_{O_i}} \perp\!\!\!\perp \mathbf{X_{O_j}} | \mathbf{X_S}$ in $\mathcal{G}$ and P. □

2.2.2 Plates

The discussion of OOGs (e.g., figure 2.2.2 in section 2.2.1) focused on attributes and relationships for an entity class ("students"). But now consider specific entities (students $p = 1, \ldots, P$), and suppose that–in addition to the objects and relationships in figure 2.2.2–there is an object O_0 that includes attributes from another entity class ("schools") that influence individual attributes for each of these students. Figure 2.2.4 illustrates the structure of the resulting OOG.

Even though figure 2.2.4 omits students $p = 2, \ldots, P-1$, the relationships are still quite complicated. Fortunately, we can take advantage of the repeated structures within this OOG and simplify these relationships using "plates" (shown as boxes around sets of objects in figure 2.2.4), which were introduced by Buntine (1994) and are used extensively in the WinBUGS graphical interface DoodleBUGS (Lunn et al., 2000). The utility of plates is that they can be "stacked" if they have repeated structures. We formalize this concept by defining two different types of stacked plates: stacked exchangeable plates, and stacked dynamic plates.

Stacked exchangeable plates

Consider an OOG $\mathcal{G}^{\mathcal{M}_n}$, and suppose that we arrange a subset of the objects in $\mathcal{G}^{\mathcal{M}_n}$ into P mutually-exclusive plates $\mathcal{P}_1, \ldots, \mathcal{P}_P$ with n_P objects in each plate, so that plate $\mathcal{P}_p$ contains objects $O_{1p}, \ldots, O_{n_P p}$ (note that n_P must be constant across plates). Property 2.2.3 describes the conditions under which plates $\mathcal{P}_1, \ldots, \mathcal{P}_P$ can be stacked as exchangeable plates.

Property 2.2.3 *Plates $\mathcal{P}_1, \ldots, \mathcal{P}_P$ can be stacked as **exchangeable plates** if and only if:*

1. *(No interference) There are no edges between an object in $\mathcal{P}_p$ and any object in another plate $\mathcal{P}_{p'} \in \{\mathcal{P}_1, \ldots, \mathcal{P}_P\}, p \neq p'$*
2. *(Repeated internal structure) Given any pair of objects $O_{ip}, O_{jp} \in \mathcal{P}_p$, $i, j \in \{1, \ldots, n_P\}$, the edge between O_{ip} and O_{jp} (if any) is the same for $p = 1, \ldots, P$.*
3. *(Consistent external structure) Given any object $O_{ip} \in \mathcal{P}_p$, $i \in \{1, \ldots, n_P\}$, and object $O_j \notin \{\mathcal{P}_1, \ldots, \mathcal{P}_P\}$, the edge between O_{ip} and O_j is the same for $p = 1, \ldots, P$*

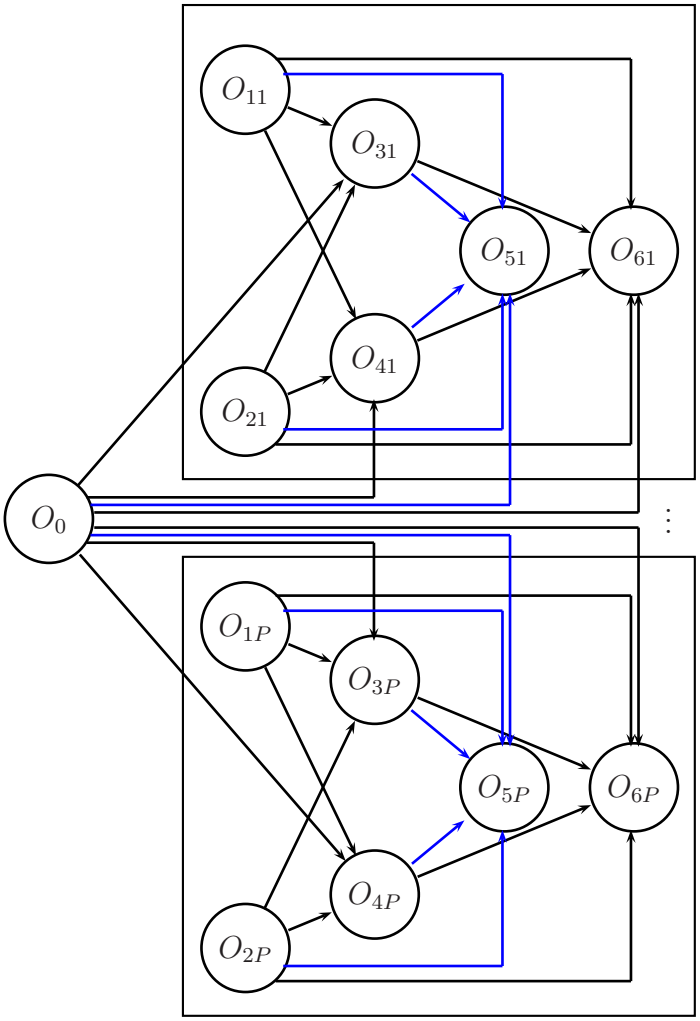


Figure 2.2.4: OOG for students (entities) $p = 1, \ldots, P$

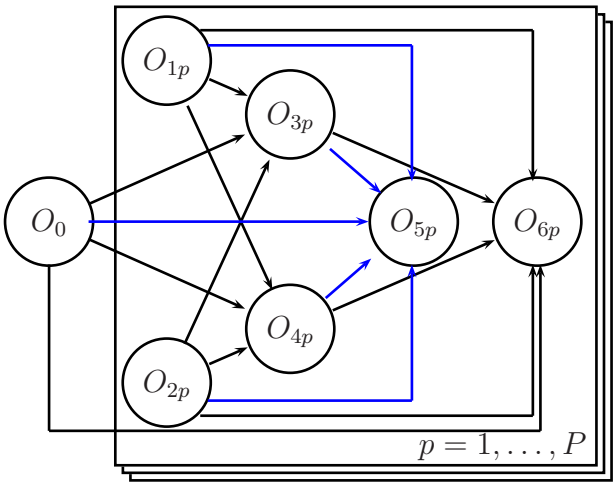


Figure 2.2.5: Stacked exchangeable plates

When plates represent collections of objects for different individuals (as in figure 2.2.4), the first part of property 2.2.3 eliminates the possibility of "interference" between individuals. The no-interference assumption SUTVA (assumption 1.2.1 from chapter 1) will not always hold in an educational intervention; for example, whether or not student 1 receives the intervention could influence the outcome of student P if student 1 is in the same classroom as student P all year (this could be represented by an edge from O_{51} to O_{6P} in figure 2.2.4: see Ogburn and VanderWeele (2014) for other graphical representations of interference between units). However, the first part of property 2.2.3 implies that SUTVA must hold.

The second and third parts of property 2.2.3 formalize the notion of "repeated structures" in an OOG. Part 2 requires that there be a repeated internal structure within each plate; i.e., that the structure of objects and edges within each plate must be identical. On the other hand, part 3 requires that there be a consistent external structure; i.e., the relationships between an object outside of the plates and objects inside the plates must be the same for all plates. It is easy to verify that property 2.2.3 holds for the OOG in figure 2.2.4. Specifically, there are no edges between the plates, the internal structure of each plate is identical, and every time there is an edge between O_0 and an object $O_{ip} \in \mathcal{P}_p$, there is an edge between O_0 and $O_{ip'}$ in each other plate $\mathcal{P}_{p'}$. By property 2.2.3, we can therefore represent the plates in figure 2.2.4 as stacked exchangeable plates, shown in figure 2.2.5 (note the index in the bottom right corner indicates that these are the stacked exchangeable plates for individuals $p = 1, \ldots, P$).

Importantly, unlike the transition from DAGs to OOGs described in section 2.2.1, the transition from the OOG in figure 2.2.4 to the stacked exchangeable plates in figure 2.2.5 involves *no loss of information*. That is, every edge and object in the OOG in figure 2.2.4 is represented in figure 2.2.5, just in a simplified format. This means that, as long as we are careful in interpreting the objects and edges in these stacked exchangeable plates (following property 2.2.3), we can apply the theorems from section 2.2.1 directly to OOGs with stacked

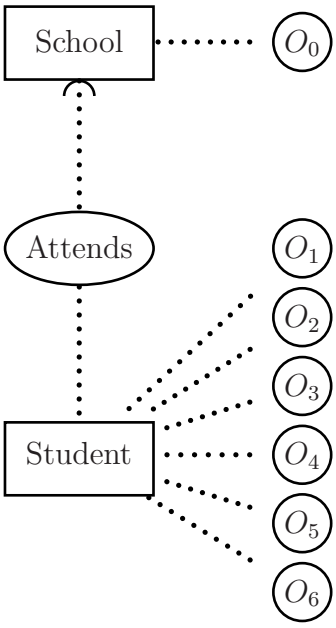


Figure 2.2.6: DAPER diagram (Heckerman et al., 2007)

exchangeable plates, like those in figure 2.2.5.

A second important aspect of the transition from figure 2.2.4 to figure 2.2.5 involves the interpretation of missing edges in stacked exchangeable plates. Specifically, property 2.2.3 ensures that an edge between objects O_{ip} and O_{jp} is missing in figure 2.2.5 if and only if the edge between $O_{ip'}$ and $O_{jp'}$ in figure 2.2.4 is missing for all $p' = 1, \ldots, P$. That is, missing edges in stacked exchangeable plates imply conditional independence *for each individual* rather than merely at the population level. The distinction is important because the absence of individual effects is not testable even in an idealized experiment. We discuss this limitation of stacked exchangeable plates further in section 2.5.

Stacked exchangeable plates are useful for nested data, such as data on students within schools. Heckerman et al. (2007) provide an alternative graphical framework for this type of data structure called Directed Acyclic Probabilistic Entity Relationship (DAPER) diagrams. Figure 2.2.6 shows a DAPER diagram for the OOG from figure 2.2.5 in a scenario in which the entity class "student" is nested within the entity class "school" (we have omitted relationships between objects for parsimony). The node "Attends" is an example of a "relationship class", and the curved arrowhead denotes a "many-to-one" relationship.

While figure 2.2.6 provides an intuitive picture of the *structure* of the data, it is difficult to represent all the *relationships* from figure 2.2.5 in this framework. Fortunately, Heckerman et al. (2007) prove that DAPERs and plate models are actually equivalent (i.e., there is a one-to-one correspondence between entity classes in the DAPER framework and plates

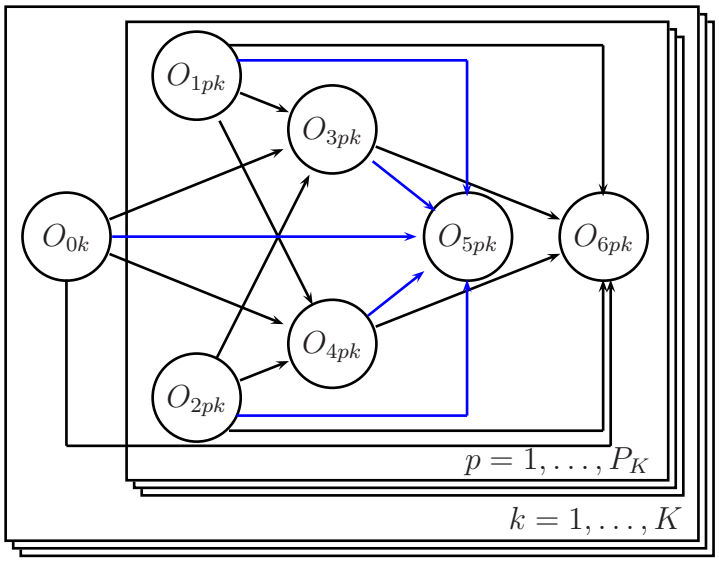


Figure 2.2.7: Stacked nested exchangeable plates

in a plate model), and concede that plates "are often easier to understand" when the data are nested. To illustrate this, figure 2.2.7 shows the same scenario using stacked nested exchangeable plates. It is straightforward to represent all relationships between objects in this framework, which allows us to apply the theorems from section 2.2.1 directly to OOGs with stacked nested exchangeable plates. On the other hand, as we discuss in section 2.5 (and as argued by Heckerman et al. (2007)), DAPER diagrams may be more useful when data are not nested.

The framework in figure 2.2.7 also allows us to incorporate "ID variables" (Rattigan and Jensen, 2010; Rattigan et al., 2011) into OOGs. An ID variable is simply an identifier for specific entities within an entity class; i.e., $I_k = k$ for $k = 1, \ldots, K$ in figure 2.2.8. Since all attributes associated with a given entity class are deterministic functions of the ID variable for that entity class, the school-level object O_{0k} becomes a deterministic object in figure 2.2.8 (i.e., all students with the same values of I_k have the same values of all the attributes in O_{0k}). The utility of ID variables will become clear when we discuss multilevel models in section 2.3.2, but we also discuss the limitations of ID variables in section 2.5 and appendix B (note that we also put a double-circle around the ID variable in figure 2.2.8 to emphasize that the ID variable and attributes associated with that entity class must be regarded as *fixed* for this framework to be consistent: see appendix B for more details).

Stacked dynamic plates

Stacked exchangeable plates provide a dramatic simplification of OOGs when there is no interference between objects in different plates. However, building on prior work with dynamic Bayesian networks (e.g. Murphy, 2002; Nodelman et al., 2002) and graphical representations of multivariate time series (e.g. Didelez, 2008; Eichler, 2007), we can extend the concept of

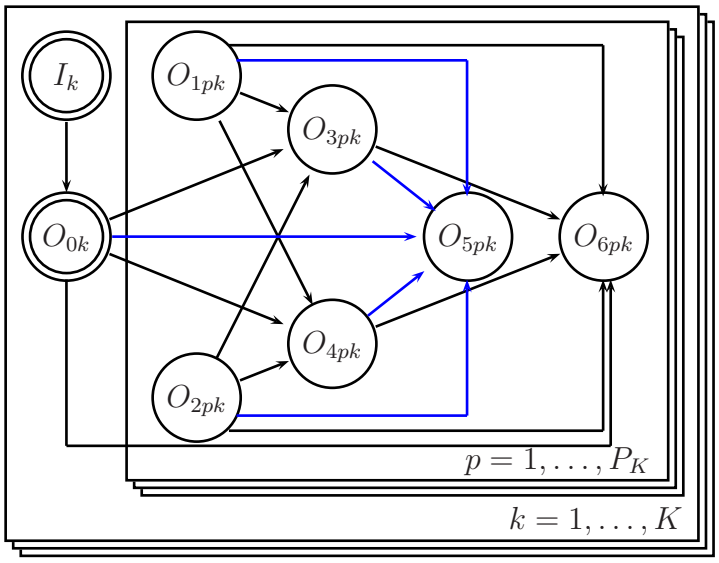


Figure 2.2.8: Stacked nested exchangeable plates with ID variable

stacked plates to allow for edges between objects in plate $\mathcal{P}_{p-1}$ and objects in plate $\mathcal{P}_p$. This is most common in a longitudinal setting when attributes within objects at time $t-1$ can influence attributes within objects at time t. As a very simple example, imagine following a single individual over multiple years. In each year $t = 1, \ldots, T$, the relationships between attributes for this individual can be represented by an OOG two objects, O_{1t} and O_{2t}. Suppose that within each year t there is an edge from O_{1t} to O_{2t}, but between years $t-1$ and t there is an edge from $O_{2(t-1)}$ to O_{1t}. Figure 2.2.9 displays the resulting OOG.

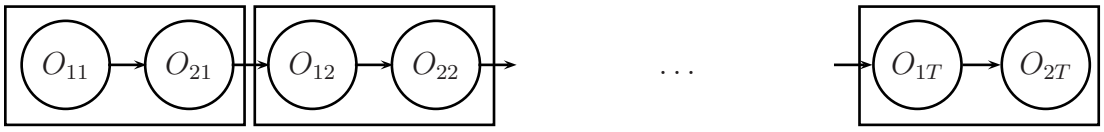


Figure 2.2.9: Dynamic network

The plates in figure 2.2.9 clearly capture a repeated structure of the OOG, but they do *not* satisfy property 2.2.3 because there is interference between plates, so they *cannot* be represented as stacked exchangeable plates. But they do satisfy a related property, property 2.2.4, that describes the conditions under which plates $\mathcal{P}_1, \ldots, \mathcal{P}_P$ can be stacked as *dynamic plates*.

Property 2.2.4 *Plates $\mathcal{P}_1, \ldots, \mathcal{P}_P$ can be stacked as **dynamic plates**, with dashed edges representing edges between plates $\mathcal{P}_{p-1}$ and $\mathcal{P}_p$, if and only if:*

1. *(Markov structure) There are no edges between an object in $\mathcal{P}_p$ and any object in*

another plate $\mathcal{P}_{p'} \in \{\mathcal{P}_1, \ldots, \mathcal{P}_P\}, p' \notin \{p-1, p, p+1\}$*, and any edge between objects in plates* $\mathcal{P}_p$ *and* $\mathcal{P}_{p+1}$ *must be directed from the object in* $\mathcal{P}_p$ *to the object in* $\mathcal{P}_{p+1}$*.*

2. *(Repeated dynamic structure) Given any pair of objects in successive plates,* $O_{i(p-1)} \in \mathcal{P}_{p-1}$ *and* $O_{jp} \in \mathcal{P}_p$ *(where* i *could equal* j*), the edge between* $O_{i(p-1)}$ *and* O_{jp} *is the same for* $p = 2, \ldots, P$

3. *(Repeated internal structure) Given any pair of objects* $O_{ip}, O_{jp} \in \mathcal{P}_p$*,* $i, j \in \{1, \ldots, n_P\}$*, the edge between* O_{ip} *and* O_{jp} *is the same for* $p = 1, \ldots, P$*.*

4. *(Consistent external structure) Given any object* $O_{ip} \in \mathcal{P}_p$*,* $i \in \{1, \ldots, n_P\}$*, and object* $O_j \notin \{\mathcal{P}_1, \ldots, \mathcal{P}_P\}$*, the edge between* O_{ip} *and* O_j *is the same for* $p = 1, \ldots, P$

Property 2.2.4 replaces the "no interference" condition of property 2.2.3 with two conditions. The first (the "Markov structure" condition) requires that edges between plates be directed from $\mathcal{P}_p$ and $\mathcal{P}_{p+1}$, while the second (the "repeated dynamic structure" condition) requires that these edges between plates be consistent across all successive plates. Since the plates in 2.2.9 follow these conditions, we can stack them as dynamic plates and, following Didelez (2008), represent edges between successive plates as dashed edges in figure 2.2.10.

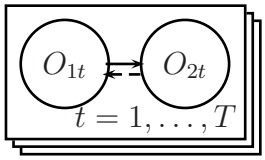


Figure 2.2.10: Stacked dynamic plates

We interpret the solid edge in figure 2.2.10 as an edge between O_{1t} and O_{2t} for $t = 1, \ldots, T$, and the dashed edge in figure 2.2.10 as an edge between $O_{2(t-1)}$ and O_{1t} for $t = 2, \ldots, T$. The utility of stacked dynamic plates will become clear when we consider complicated longitudinal settings in section 2.3.3. For now, it is worth stressing that, just like with stacked exchangeable plates, there is no information lost in the transition from the OOG in figure 2.2.9 to the OOG with stacked dynamic plates in figure 2.2.10; every object and edge from figure 2.2.9 is still represented in figure 2.2.10. Therefore, as long as we are careful about tracing paths across different plates (using the rules in property 2.2.4), we can also directly apply the theorems from section 2.2.1 to OOGs with stacked dynamic plates.

2.2.3 Node-splitting

As we discuss in section 1.2.3 of chapter 1, Richardson and Robins (2013) recently introduced Single World Intervention Templates (SWITs), a unification of the graphical and potential outcomes approaches to causality. Specifically, Richardson and Robins (2013) develop a "node-splitting" operation on intervention nodes in DAGs that permits the inclusion of potential outcomes in a graphical framework. Thus as the last step in our development in this

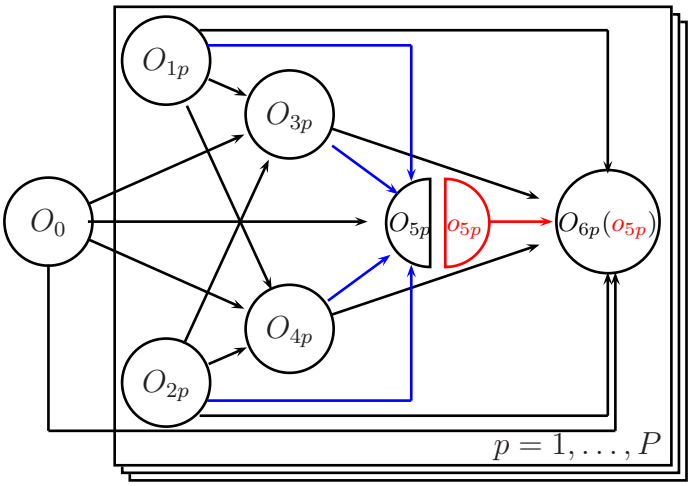


Figure 2.2.11: Single-World Object-Oriented Plates (SWOOPs)

section, we apply the same node-splitting operation to OOGs with plates and derive Single World Object-Oriented Plates (SWOOPs).

We return to the OOG with stacked exchangeable plates shown in figure 2.2.5. Recall that in this OOG, O_0 is a complex object containing school-level attributes, while $O_1, \ldots, O_4$ are complex objects containing individual-level attributes that are observed and time-invariant, unobserved and time-invariant, observed and time-variant, and unobserved and time-variant, respectively. On the other hand, O_5 and O_6 are both simple objects: O_5 contains the intervention attribute (the "intervention object"), while O_6 contains the outcome attribute (the "outcome object"). Following the same procedure described in section 1.2.3 of chapter 1, we can split the intervention object into the random variable O_5 and the fixed variable o_5, and relabel the outcome object as a potential outcome $O_6(o_5)$ (Neyman, 1923; Rubin, 1974) that depends on the value of the intervention (o_5); note that node-splitting represents an additional causal assumption beyond the Markov property represented by the original graph. Figure 2.2.11 shows the resulting SWOOPs.

For a given value of o_{5p}, the SWOOPs in figure 2.2.11 are derived from SWITs using the exact same process described in section 2.2.1 for deriving OOGs from DAGs. We can therefore apply the theorems from section 2.2.1 to SWOOPs like the one in figure 2.2.11 and establish d-separation relationships between sets of attributes in underlying SWITs. This in turn will allow us to use SWOOPs to derive *counterfactual* independence relationships (since there is a one-to-one relationship between the intervention and outcome objects and the attributes they contain, we use the convention $\mathbf{X}_{\mathbf{O_i}} \equiv O_i$ for these objects in these derivations). We discuss this application of SWOOPs extensively in the next section.

2.3 Inference

The development in section 2.2 assumed that the underlying DAG is completely known (i.e., the DAG contains all the relevant attributes, and the edges in this DAG represent known relationships between these attributes). But SWOOPs are particularly useful when it is not possible to completely represent the structure of the underlying DAG, and/or when the relationships between some attributes in the underlying DAG are unknown. These scenarios are particularly common in the multivariate, multilevel, and longitudinal settings common in education research. In these settings, SWOOPs offer a "shorthand" for representing relationships between *groups* of attributes, and the theorems from section 2.2 allow us to infer conditional independence relationships in the (potentially incomplete and/or unknown) underlying DAG and draw causal conclusions. In this section, we illustrate the utility of SWOOPs for causal inference in multivariate settings (section 2.3.1), multilevel settings (section 2.3.2), and longitudinal settings (section 2.3.3).

2.3.1 Multivariate Settings

Consider the following simple example of a non-randomized educational intervention: students $p = 1, \ldots, P_K$ in a single school $k = 1$ and year $t = 1$ are targeted for an intervention on the basis of attributes that may or may not be observable by researchers. Students selected for the intervention ($A_{pkt} = 1$) form the treatment group, while students not selected for the intervention ($A_{pkt} = 0$) form the control group (note that the k and t subscripts, while useful in subsequent examples, are not necessary for this example since $k = 1$ and $t = 1$ for all students). At the end of the year, an observed outcome Y_{pkt} is measured for students in both the treatment group and control group. As in section 1.2.1 of chapter 1, we can define two potential outcomes for each student p: $Y_{pkt}(a_{pkt} = 1) \equiv Y_{pkt}(1)$ is the outcome for student p if (potentially contrary to fact) student p received the intervention; and $Y_{pkt}(a_{pkt} = 0) \equiv Y_{pkt}(0)$ is the outcome for student p if (potentially contrary to fact) student p did not receive the intervention. Of course, only one potential outcome is observed for each student: $Y_{pkt}(1)$ if student p is in the treatment group; or $Y_{pkt}(0)$ if student p is in the control group.

The goal in evaluating an intervention like this is often to estimate the average treatment effect (ATE) $\mathrm{E}(Y_{pkt}(1) - Y_{pkt}(0))$; that is, the average difference between the outcome students *would have* achieved if they had received the intervention and the outcome students *would have* achieved if they had not received the intervention. A common approach to estimating the ATE is to use a covariate adjustment model (e.g., linear regression, propensity scores, etc.) that "controls for" a vector $\mathbf{X_{pkt}}$ of observed attributes for each student p (note that we will consider specific examples of these models in section 2.4). Under the homogeneity, linearity, and no-interaction assumptions discussed in section 1.2.1 of chapter 1, these models provide an unbiased estimate of the ATE under condition 2.3.1 (Angrist and Pischke, 2008; Rosenbaum and Rubin, 1983):

Condition 2.3.1 $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(1) \mid \mathbf{X_{pkt}}$ *and* $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(0) \mid \mathbf{X_{pkt}} \quad \forall\, p, p', k, k', t, t'.$

Condition 2.3.1 is actually stronger than necessary, because it implies independence for *all*

individuals rather than just independence at the population level. We use condition 2.3.1 because, as we discuss in section 2.2.2, the absence of edges in stacked SWOOPs imply the absence of causal effects at the individual level, but we discuss the drawbacks of this decision in section 2.5.

The key question in this section, then, is under what assumptions does condition 2.3.1 hold? In the previous chapter, we demonstrated the utility of graphs (and SWIGs in particular) for communicating these assumptions and verifying counterfactual independence relationships like condition 2.3.1. But it is practically impossible to represent even this relatively simple example in a traditional graph for two reasons, both of which are related to the *multivariate* nature of the data. First, although researchers know which observed attributes to include in the graph (i.e., the attributes in $\mathbf{X_{pkt}}$), this vector potentially includes a large number of attributes, and the relationships *between* these attributes may be unknown. Second, it may not be possible to enumerate all of the *unobserved* attributes that could confound the relationship between the intervention and outcome in the study.

We therefore use this example as our first illustration of how SWOOPs can be used to infer counterfactual independence relationships like condition 2.3.1 that validate causal conclusions from observational data. Specifically, instead of attempting to enumerate all the variables in this example, define four groups of variables (note that these are analogous to the objects defined in section 2.2.1):

- $\mathbf{O_{pk}}$ is the collection of attributes for student p in school k that are *time-invariant* and *observed* to researchers
- $\mathbf{O_{pkt}}$ is the collection of attributes for student p in school k and year t that are *time-variant* and *observed* to researchers
- $\mathbf{U_{pk}}$ is the collection of attributes for student p in school k that are *time-invariant* and *unobserved* to researchers
- $\mathbf{U_{pkt}}$ is the collection of attributes for student p in school k and year t that are *time-variant* and *unobserved* to researchers

We can communicate assumptions about the relationships between these four collections of attributes, the intervention, and the outcome using the SWOOPs $\mathcal{G}^{\mathcal{M}_6}$ in figure 2.3.1. We note that some of the conventions in figure 2.3.1 are similar to the conventions in chapter 1, but some are different: black nodes represent collections of variables that are observed to researchers; gray nodes represent collections of variables that are unobserved to researchers; black edges represent edges that are assumed to exist; blue edges represent edges that *might* exist under different assumptions; and the red edge represents the treatment effect of interest. Note that relationships for students $p = 1, \ldots, P_K$ are represented by stacked exchangeable plates, while the other plates in figure 2.3.1 (the plate for time $t = 1$ and school $k = 1$) are extraneous for this example because it involves only a single school in a single year. We will expand this example to multiple schools and years in sections 2.3.2 and 2.3.3, respectively.

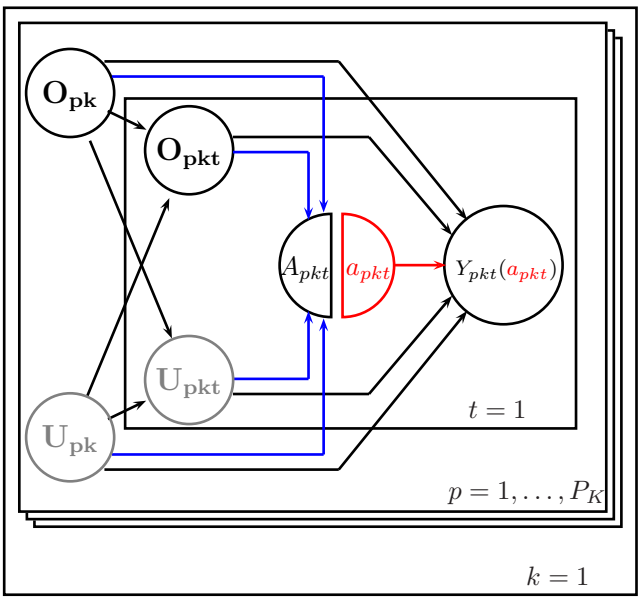


Figure 2.3.1: Stacked SWOOPs in multivariate setting (1 school, 1 year)

We also pause to note a shift from the development in section 2.2. In that section, we assumed knowledge of the relationships in the underlying SWIT $\mathcal{G}$. But the SWOOPs $\mathcal{G}^{\mathcal{M}_6}$ in figure 2.3.1 could have been derived from a number of different underlying SWITs $\mathcal{G}$, each of which has different relationships between attributes *within* the objects in $\mathcal{G}^{\mathcal{M}_6}$ and different numbers of edges *between* attributes in different objects in $\mathcal{G}^{\mathcal{M}_6}$. The edges in $\mathcal{G}^{\mathcal{M}_6}$ represent broad assumptions that unite these different underlying SWITs. For example, the edge between $\mathbf{O_{pk}}$ and $Y_{pkt}(a_{pkt})$ represents the assumption that *at least one* observed, time-invariant attribute (e.g., gender) influences the outcome. Likewise, the edge between $\mathbf{U_{pk}}$ and $\mathbf{O_{pkt}}$ represents the assumption that *at least one* unobserved, time-invariant attribute (e.g., IQ) influences *at least one* time-variant, observed attribute (e.g., special education classification). These assumptions will vary from setting to setting (and potentially from researcher to researcher!), but the relationships shown in figure 2.3.1 represent the assumptions we might make in a "typical" non-randomized educational intervention.

The most important edges in figure 2.3.1 are the four blue edges between objects representing different collections of attributes and the intervention node A_{pkt}. The *absence* of these edges represent broad assumptions about how students are selected for the intervention. For example, under the assumption that students are not selected for the intervention on the basis of any unobserved, time-variant characteristics, the edge between $\mathbf{U_{pkt}}$ and A_{pkt} is missing. These types of assumptions (e.g., "no unmeasured confounding") are central to establishing the counterfactual independencies (such as condition 2.3.1) that justify causal conclusions from observational data.

For example, consider the assumptions represented by the SWOOPs in figure 2.3.2. The missing edges between $\mathbf{U_{pk}}$ and A_{pkt} and between $\mathbf{U_{pkt}}$ and A_{pkt} in this figure encode the assumption that students are not selected for the intervention on the basis of any unob-

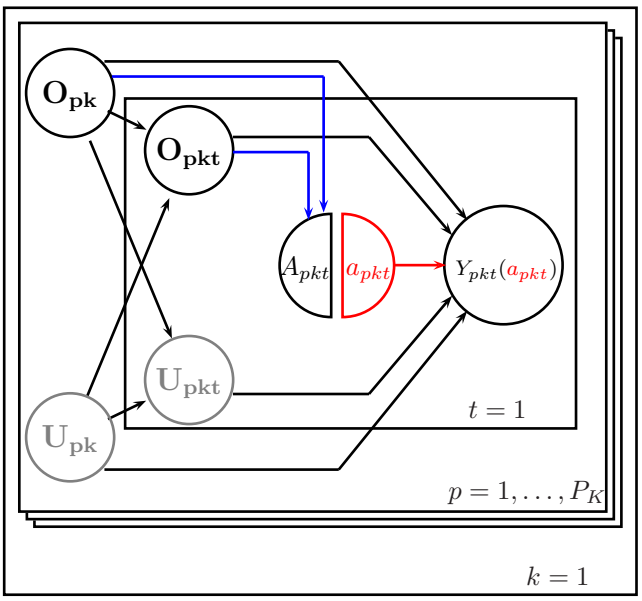


Figure 2.3.2: Conditions justifying covariate adjustment model in multivariate setting

served characteristics, either time-variant or time-invariant. It is well-known (e.g. Angrist and Pischke, 2008) that this "selection on observables" assumption is a necessary condition for covariate adjustment models to produce an unbiased estimate of the ATE, and we can formalize this by using figure 2.3.2 to connect this assumption to the counterfactual independence condition 2.3.1. Specifically, every path between A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ in figure 2.3.2 contains the non-collider $\mathbf{O_{pk}}$ or $\mathbf{O_{pkt}}$. Thus A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ are d-separated given the set $S = \{\mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$ $\forall\ p, p', k, k', t, t'$ in figure 2.3.2. We can therefore apply theorem 2.2.2 from section 2.2.1: for any $a_{p'k't'}$, $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(a_{p'k't'}) \mid \mathbf{X}_{\{\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$ $\forall\ p, p', k, k', t, t'$ in any distribution P that is Markov with respect to the underlying SWIT $\mathcal{G}$ (we will omit this qualification in future sections and just assume that the Markov property holds in any underlying SWIT). So, as long as the covariate adjustment model controls for all the observed attributes in $\mathbf{X}_{\{\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$, condition 2.3.1 holds under the assumptions shown in figure 2.3.2 and causal conclusions from the model are justified. Again, we stress that the SWOOPs in figure 2.3.2 allow us to draw this conclusion without complete knowledge of the structure of the underlying SWIT $\mathcal{G}$.

2.3.2 Multilevel Settings

We now extend the simple example in section 2.3.1 to an intervention that is implemented across multiple schools $k = 1, \ldots, K$. This example will allow us to extend the results from section 2.3.1 to models that include either school-level covariates or a "school effect" (e.g., in a multilevel model: see Gelman and Hill (2006)). The relationships in the stacked independent SWOOPs in figure 2.3.3 represent what we might expect in a "typical" non-randomized educational intervention across multiple schools.

Figure 2.3.3 includes three new objects in addition to the objects discussed in figures

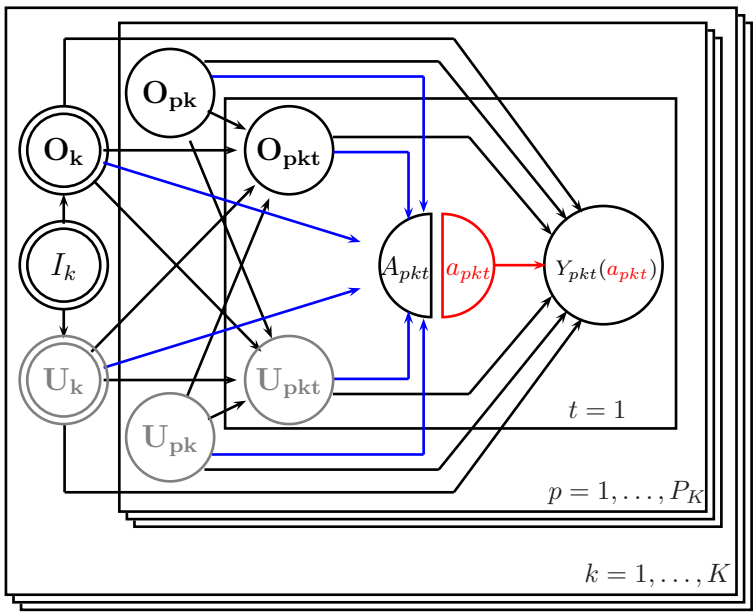


Figure 2.3.3: Stacked SWOOPs in multilevel setting (K schools)

2.3.1 and 2.3.2:

- $\mathbf{O_k}$ is the collection of attributes for school k that are *observed* to researchers
- $\mathbf{U_k}$ is the collection of attributes for school k that are *unobserved* to researchers
- I_k is an ID object for school k (i.e., $I_k = k$ for $k = 1, \ldots, K$)

Note that $\mathbf{O_k}$ and $\mathbf{U_k}$ are both deterministic objects because, by definition, all school-level attributes are a deterministic function of the ID variable I_k.

In a multilevel intervention, school-level information can be used in (at least) two different ways to control for potential school-level confounding and estimate the ATE $\mathrm{E}(Y_{pkt}(1) - Y_{pkt}(0))$. The first option is to include observed school-level attributes in the vector of control variables $\mathbf{X}_{\{\mathbf{O_k},\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$ and estimate a covariate adjustment model just like in section 2.3.1. The SWOOPs in figure 2.3.4 illustrate the assumptions that justify this approach. As before, these assumptions can be characterized as assuming that selection for the intervention is based solely on observed attributes (i.e., all edges from objects containing unobserved attributes and the intervention node are missing in figure 2.3.4).

As before, we can use figure 2.3.4 to connect these assumptions to condition 2.3.1, but we have to use the theorem from section 2.2.1 (theorem 2.2.5) that allows for deterministic objects. Specifically, since every path between A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ in figure 2.3.4 contains a non-collider in the set $S = \{\mathbf{O_k}, \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$, A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ are D-separated given the set S $\forall\ p, p', k, k', t, t'$ in figure 2.3.4 by definition 2.2.9. This allows us to apply theorem 2.2.3: for any $a_{p'k't'}$, $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(a_{p'k't'}) \mid \mathbf{X}_{\{\mathbf{O_k},\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$ $\forall\ p, p', k, k', t, t'$. Therefore, controlling for all the observed attributes in $\mathbf{X}_{\{\mathbf{O_k},\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$ is sufficient to ensure that condition 2.3.1 holds under the assumptions shown in figure 2.3.4 and causal conclusions from

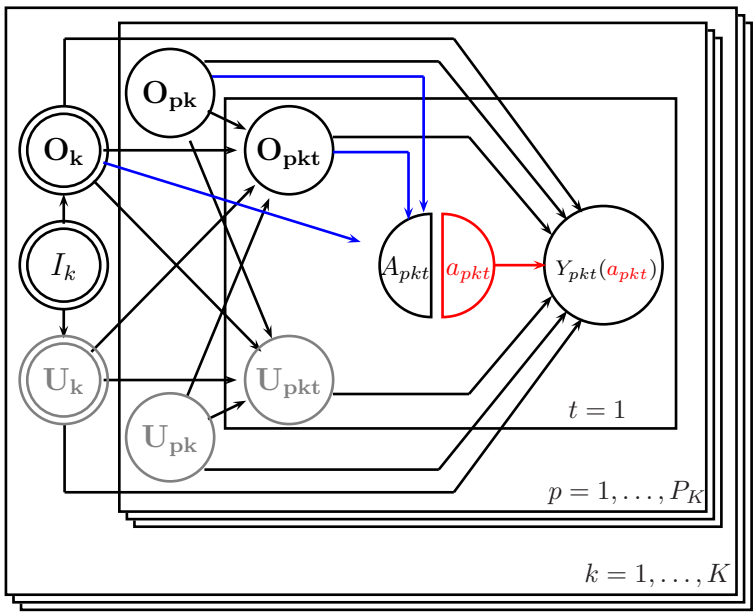


Figure 2.3.4: Conditions justifying covariate adjustment model in multilevel setting

a covariate adjustment model are justified.

A second option with multilevel data, though, is to estimate a multilevel model (e.g., a Hierarchical Linear Model (HLM, Raudenbush and Bryk, 2002)) that includes an effect for each individual school. By controlling for all variation at the school-level (and under linearity and no-interaction assumptions), these models can control for confounding by *any* school-level attribute, observed or otherwise. This is shown in the stacked independent SWOOPs in figure 2.3.5.

The edge in figure 2.3.5 that does not appear in figure 2.3.4 is the edge between $\mathbf{U_k}$ and A_{pkt}; i.e., the assumptions represented in figure 2.3.5 allow for the possibility that students are selected for the intervention in part because of *unobserved* attributes in school k. Even under these weaker assumptions, we can use figure 2.3.5 to verify condition 2.3.1 when the vector of control variables includes the school indicator I_k (i.e., so the vector of control variables is $\mathbf{X}_{\{\mathbf{I_k},\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$, where $\mathbf{X}_{\mathbf{I_k}}$ is a set of indicator variables for each school). To do this, note that every path between A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ in figure 2.3.5 contains a non-collider that is functionally determined (see definition 2.2.8) by the set $S = \{I_k, \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$. Therefore, A_{pkt} and $Y_{p'k't'}(a_{p'k't'})$ are D-separated given the set S $\forall$ p, p', k, k', t, t' in figure 2.3.5 by definition 2.2.9. Thus by theorem 2.2.3, for any $a_{p'k't'}$, $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(a_{p'k't'}) \mid \mathbf{X}_{\{\mathbf{I_k},\mathbf{O_{pk}},\mathbf{O_{pkt}}\}}$ $\forall$ p, p', k, k', t, t', meaning that causal conclusions from a multilevel model are justified under the assumptions shown in figure 2.3.5. This validates the use of multilevel models when there is school-level variation in the selection of students for a cross-school intervention.

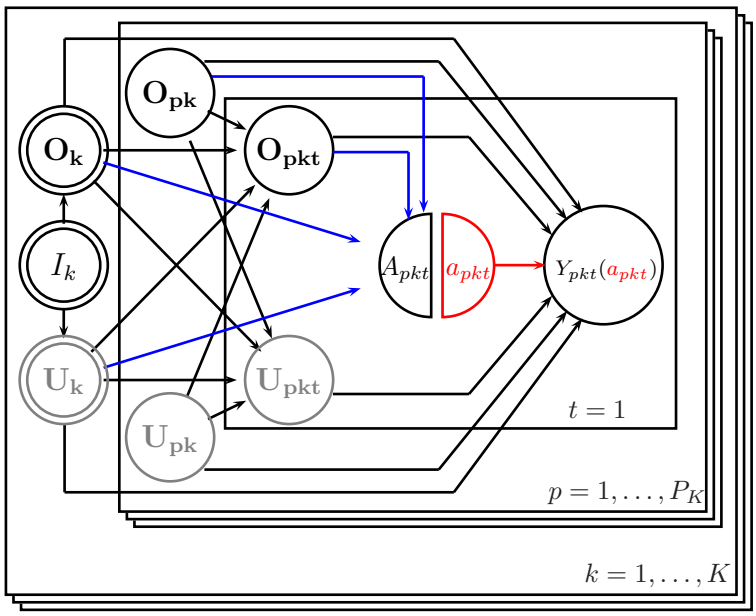


Figure 2.3.5: Conditions justifying model with school effect in multilevel setting

2.3.3 Longitudinal Settings

We now extend the simple example from section 2.3.1 to an intervention that is implemented across multiple years $t = 1, \ldots, T$. We begin by considering just school $k = 1$ and assuming that students do not switch schools between years in the study. In this longitudinal setting, we can consider both covariate adjustment models and models with "student effects" (which are identified by variation within students across the different years of the study). These examples are instructive because, unlike in section 2.3.2 where the assumptions that justify a multilevel model are a subset of the assumptions that justify a covariate adjustment model (suggesting, not surprisingly, that multilevel models are always preferable in this specific multilevel setting), we demonstrate in this section that covariate adjustment models and models with a student effect are preferable in a longitudinal setting under different (and non-overlapping) sets of assumptions.

The stacked SWOOPs in figure 2.3.6 include three changes from the SWOOPs in figure 2.3.1. First, while the student plates are still stacked as exchangeable plates across students $p = 1, \ldots, P$, the time plates are stacked dynamic plates that include a number of relationships between objects in time $t - 1$ and objects in time t (shown as dashed lines in figure 2.3.1). For example, the dashed directed edge from $\mathbf{O_{pkt}}$ to $\mathbf{U_{pkt}}$ implies that at least one time-variant, observed attribute in year $t - 1$ is assumed to influence at least one time-variant, unobserved attribute in year t (note that the two dashed lines between $\mathbf{O_{pkt}}$ and $\mathbf{U_{pkt}}$ do *not* create a cycle because there are no edges from year t back to year $t - 1$).

Second, the outcome object has been re-labeled as $Y_{pkt}(\overline{\mathbf{a}}_{\mathbf{pkt}})$, where $\overline{\mathbf{a}}_{\mathbf{pkt'}} = \{a_{pk1}, \ldots, a_{pkt'}\}$ for $t' = 1, \ldots, T$. This reflects the reality (as shown by the dashed edge representing the edge between $Y_{pk(t-1)}(\overline{\mathbf{a}}_{\mathbf{pk(t-1)}})$ and $Y_{pkt}(\overline{\mathbf{a}}_{\mathbf{pkt}})$) that a student's outcome in year t depends on the

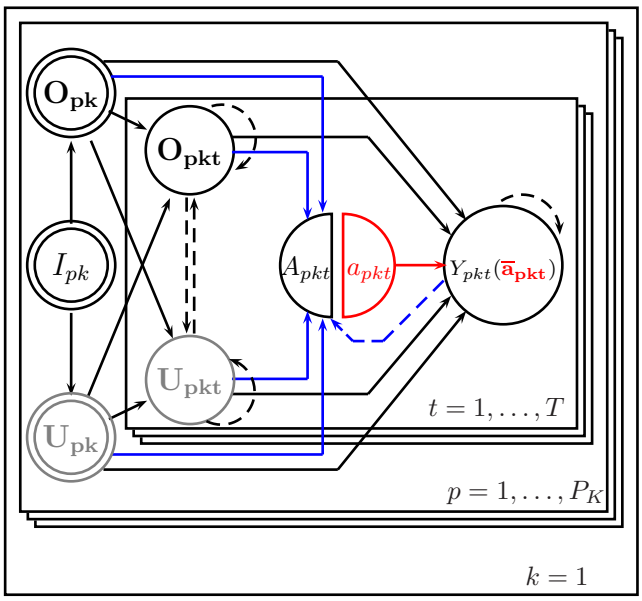


Figure 2.3.6: Stacked SWOOPs in longitudinal setting (T years)

value of all the interventions they've received in previous years. We also note the presence of the dashed blue edge representing a directed from $Y_{pk(t-1)}(\overline{\mathbf{a}}_{\mathbf{pk(t-1)}})$ to A_{pkt}, which represents the possibility that students could be selected for the intervention in year t on the basis of their outcome in year $t-1$.

Finally, just as we were able to introduce a school ID variable in the multilevel setting, we are now able to include a student ID variable I_{pk} that is unique for each student p in school k. Note that once we introduce I_{pk}, the objects $\mathbf{O_{pk}}$ and $\mathbf{U_{pk}}$ become deterministic objects, because (by definition) the time-invariant attributes that are mapped to these objects do not vary within students.

We can now use these SWOOPs to illustrate assumptions that justify causal conclusions from different models in a longitudinal setting. The SWOOPs in figure 2.3.7 represent assumptions that justify a covariate adjustment model that controls for all observed attributes in year t *and* the lagged outcome from year $t-1$. The missing edges between $\mathbf{U_{pk}}$ and A_{pkt} and between $\mathbf{U_{pkt}}$ and A_{pkt} in this figure encode the assumption that students are not selected for the intervention on the basis of any unobserved characteristics, either time-variant or time-invariant. Note that every path between every path between $A_{pk(t-1)}$ and $Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{pkt}})$ in figure 2.3.7 contains the non-collider $Y_{pk(t-1)}(\overline{\mathbf{a}}_{\mathbf{pk(t-1)}})$, while every path between A_{pkt} and $Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{pkt}})$ in figure 2.3.7 contains a non-collider in the set $\{\mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$. Therefore A_{pkt} and $Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{p'k't'}})$ are D-separated given the set $\{Y_{pk(t-1)}(\overline{\mathbf{a}}_{\mathbf{pk(t-1)}}), \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$ $\forall\ p, p', k, k', t, t'$ in figure 2.3.4 by definition 2.2.9. Thus we can apply theorem 2.2.3 and conclude that, for all sets $\overline{\mathbf{a}}_{\mathbf{pkt}}$ and $\forall\ p, p', k, k', t, t'$, $A_{pkt'} \perp\!\!\!\perp Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{p'k't'}}) \mid \mathbf{X}_{\{\mathbf{Y_{pk(t-1)}}(\overline{\mathbf{a}}_{\mathbf{pk(t-1)}}), \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}}$. Therefore, a covariate adjustment model that controls for the lagged outcome is sufficient to ensure that condition 2.3.1 holds under the assumptions shown in figure 2.3.4; namely, that there *may be* "dynamic selection" (i.e., selection for the intervention in year t based on

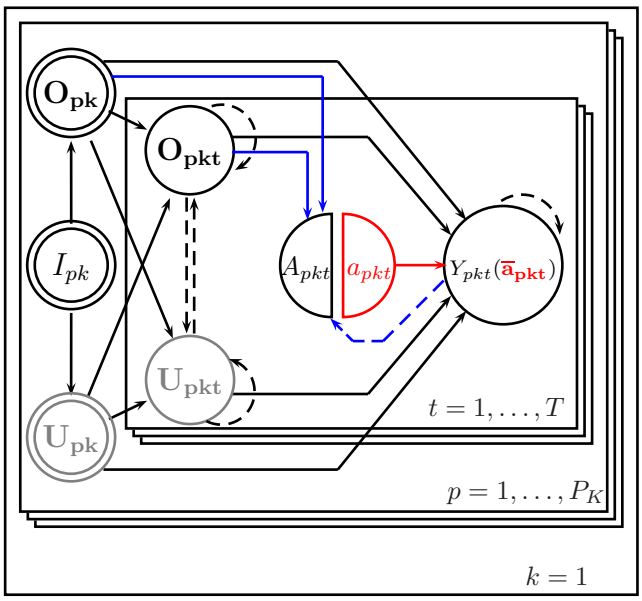


Figure 2.3.7: Conditions justifying covariate adjustment model in longitudinal setting

outcomes in year $t-1$), but that there is *no* selection based on unobserved attributes.

On the other hand, the SWOOPs in figure 2.3.8 represent assumptions that justify a model with an effect for each student that is pooled across years. The missing solid edge between $\mathbf{U_{pkt}}$ and A_{pkt} and the missing dashed edge between $Y_{pkt}(\overline{\mathbf{a}}_{\mathbf{pkt}})$ and A_{pkt} in this figure encode the assumption that students are not selected for the intervention on the basis of unobserved, time-variant characteristics or on the basis of observed prior outcomes. Note that every path between A_{pkt} and $Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{p'k't'}})$ in figure 2.3.8 contains a non-collider that is functionally determined by the set $S = \{I_{pk}, \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}$. Therefore, A_{pkt} and $Y_{p'k't'}(\overline{\mathbf{a}}_{\mathbf{p'k't'}})$ are D-separated given the set $S \;\; \forall\, p, p', k, k', t, t'$ in figure 2.3.5 by definition 2.2.9. Thus by theorem 2.2.3, for all sets $\overline{\mathbf{a}}_{\mathbf{pkt}}$ and $\forall\, p, p', k, k', t, t',\; A_{pkt'} \perp\!\!\!\perp Y_{pkt}(\overline{\mathbf{a}}_{\mathbf{pkt}})) \mid \mathbf{X}_{\{\mathbf{I_{pk}}, \mathbf{O_{pk}}, \mathbf{O_{pkt}}\}}$, meaning that causal conclusions from a model with student effects are justified under the assumptions shown in figure 2.3.8; namely, that there *may be* sorting based on unobserved attributes that do not change over time (represented by the blue edge from $\mathbf{U_{pk}}$ to A_{pkt}), but there is *not* any dynamic selection.

We conclude this section by illustrating that we can combine the stacked SWOOPs from sections 2.3.2 and 2.3.3 to permit causal reasoning in a setting with K schools and T years. We illustrate these SWOOPs in figure 2.3.9, and note that these SWOOPs allow us to investigate models with both student and school effects in a scenario where students do not switch schools over the years of the study (we discuss the limitations of this framework in section 2.5).

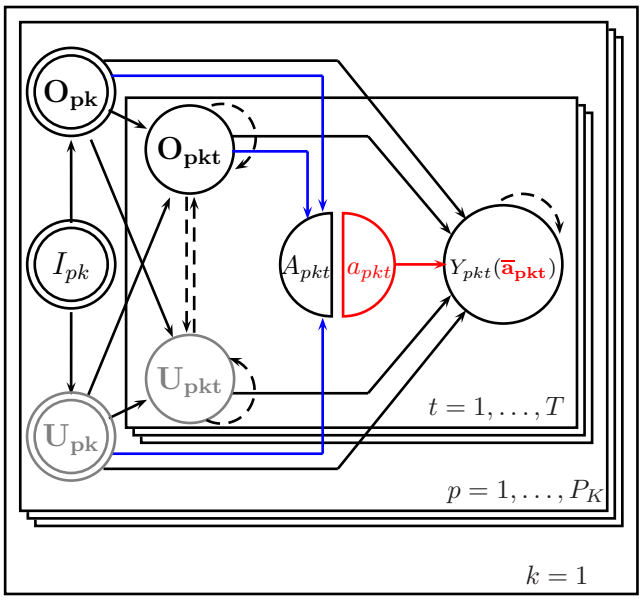


Figure 2.3.8: Conditions justifying model with student effect in longitudinal setting

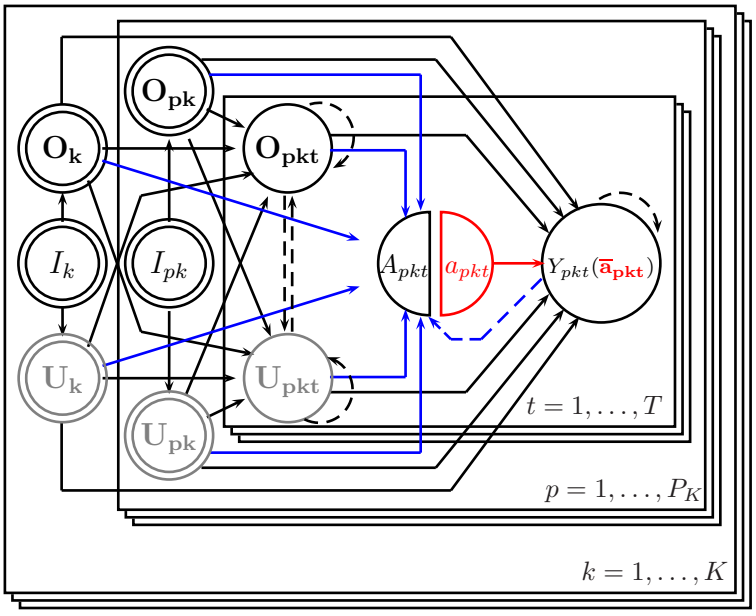


Figure 2.3.9: Stacked SWOOPs in multilevel and longitudinal setting (K schools and T years)

2.4 Application to Value-Added Models (VAMs)

Teacher evaluation provides a particularly good, and high-profile, example of the challenges that can arise from observational data. Districts and states–spurred by evidence that teachers are the most important schooling resource influencing student achievement (Chetty et al., 2013; Rivkin et al., 2005; Rockoff, 2004)–are increasingly using student test scores to evaluate teachers (Goldhaber and Theobald, 2013a). A large literature (e.g. McCaffrey et al., 2004) describes and estimates value-added models (VAMs) designed to isolate the contribution of individual teachers to student test achievement. However, outside of a small number of experiments (Glazerman et al., 2013; Kane and Staiger, 2008; Kane et al., 2013), students are not typically randomly assigned to teachers in U.S. public schools. This means that students in different classrooms may differ in a number of important ways *other than* their teacher, which raises the possibility of confounding; that is, that estimates of teacher effectiveness from VAMs are capturing the impact of these other variables rather than true teacher effects. Given this reality, it is essential to communicate and verify the conditions that justify causal conclusions from VAMs.

An emerging literature (e.g., Reardon and Raudenbush, 2009; Rothstein, 2009) seeks to describe the conditions that justify causal conclusions from VAMs. These papers provide an interesting case study in that they describe these conditions in very different terms. Rothstein (2009) describes **substantive conditions** under which VAMs may or may not produce unbiased estimates of true teacher effects; e.g., estimates from VAMs are biased if principals systematically assign students to teachers on the basis of characteristics that are unobserved to researchers and that influence student test performance. Reardon and Raudenbush (2009), on the other hand, describe **counterfactual conditions** within a potential outcomes framework; e.g., estimates from VAMs are unbiased if a student's potential outcomes are conditionally independent of teacher assignment given observable student characteristics. However, it is not immediately clear how the conditions described in these papers relate to each other, and perhaps more importantly, how researchers can communicate and verify these conditions in a real-life setting.

In this section, we use the SWOOPs from section 2.3 to demonstrate that the substantive conditions in Rothstein (2009) are equivalent the counterfactual conditions in Reardon and Raudenbush (2009). We first adapt our notation from section 2.3 to a simple example of a value-added teacher evaluation for a single grade level in a single school ($k = 1$) and year ($t = 1$). Suppose there are A teachers to whom students could be assigned in this grade, school, and year, and define the following variables for each student p (we maintain the k and t subscripts so we can expand this example to multiple schools and years):

Definition 2.4.1 *$A_{pkt} = a$ if student p is assigned to Teacher a ($a = 1, \ldots, A$)*

Definition 2.4.2 *Y_{pkt} = student p's score on the test at the end of year t*

In this setting, we can define A potential outcomes for each student p: $Y_{pkt}(a)$ is student p's test score if (potentially contrary to fact) he or she was in Teacher a's classroom in year t. As discussed in section 1.2.1 of chapter 1, the "fundamental problem of causal inference"

(Holland, 1986) is that we can only observe one of these potential outcomes for each student (i.e., the potential outcome that corresponds to the student's *actual* teacher in year t).

The goal of a value-added model (VAM) in this setting is to estimate the average treatment effect on student test performance of each teacher in the school relative to every other teacher in the school. This is often done by first designating a "reference teacher" (we will make this Teacher 1) and estimating the average treatment effect for each Teacher a' relative to Teacher 1, $a' = 2, \ldots, A$. Each of these average treatment effects can be expressed in terms of potential outcomes as $\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1)) = \mathrm{E}(Y_{pkt}(a')) - \mathrm{E}(Y_{pkt}(1))$; that is, the expected difference between how students *would have scored* if they had been in Teacher a''s classroom in year t and how students *would have scored* if they had been in Teacher 1's classroom in year t. Note that the average treatment effects $\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1)),\ a' = 2, \ldots, A$ can be used to derive the average treatment effect of any teacher a'' relative to any other teacher a'''. For example, if $a'', a''' \in \{2, \ldots, A\}$. we can derive the average treatment effect of teacher a'' relative to teacher a''' as follows:

$$\begin{aligned}
\mathrm{E}(Y_{pkt}(a'') - Y_{pkt}(a''')) &= \mathrm{E}(Y_{pkt}(a'')) - \mathrm{E}(Y_{pkt}(a''')) \\
&= \mathrm{E}(Y_{pkt}(a'')) - \mathrm{E}(Y_{pkt}(1)) - \mathrm{E}(Y_{pkt}(a''')) + \mathrm{E}(Y_{pkt}(1)) \\
&= \{\mathrm{E}(Y_{pkt}(a'')) - \mathrm{E}(Y_{pkt}(1))\} - \{\mathrm{E}(Y_{pkt}(a''')) - \mathrm{E}(Y_{pkt}(1))\} \\
&= \mathrm{E}(Y_{pkt}(a'') - Y_{pkt}(1)) - \mathrm{E}(Y_{pkt}(a''') - Y_{pkt}(1)).
\end{aligned}$$

Thus in this example, it is sufficient to estimate $A-1$ average treatment effects: $\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1)),\ a' = 2, \ldots, A$. To estimate these treatment effects, researchers typically estimate some variant of the following VAM that controls for observable student characteristics $\mathbf{X_{pkt}}$ (such as a lagged test score and various demographic characteristics) and includes a fixed effect for each teacher a, $a = 2, \ldots, A$ (represented by the indicators $\{I_{a=a'}\}$, where $\{I_{a=a'}\} = 1$ if $a = a'$ and $\{I_{a=a'}\} = 0$ otherwise):

$$\mathrm{E}(Y_{pkt}|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = a) = \alpha_0 + \boldsymbol{\alpha}_1^T \mathbf{x_{pkt}} + \alpha_2\{I_{a=2}\} + \ldots + \alpha_A\{I_{a=A}\}. \qquad (2.4.1)$$

But under what conditions are the estimated regression coefficients $\hat{\alpha}_{a'}$ from equation 2.4.1 unbiased estimates of the average treatment effects $\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1))$? Under linearity and homogeneity assumptions described in Reardon and Raudenbush (2009), each linear regression coefficient $\alpha_{a'}$ $(a' = 2, \ldots, A)$ in equation 2.4.1 is the expected difference (over values $\mathbf{x_{pkt}}$ of $\mathbf{X_{pkt}}$) in the expected value of Y_{pkt} between students with $\mathbf{X_{pkt}} = \mathbf{x_{pkt}}$ in Teacher a''s classroom ($A_{pkt} = a'$) versus students with $\mathbf{X_{pkt}} = \mathbf{x_{pkt}}$ in Teacher 1's classroom ($A_{pkt} = 1$):

$$\alpha_{a'} = \mathrm{E}\{\mathrm{E}(Y_{pkt}|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = a') - \mathrm{E}(Y_{pkt}|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = 1)\}. \qquad (2.4.2)$$

Reardon and Raudenbush (2009) show the estimand in equation 2.4.2 to equal the average treatment effect $\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1))$ under the following **counterfactual condition**:

Condition 2.4.1 $A_{pkt} \perp\!\!\!\perp Y_{pkt}(a) \mid \mathbf{X_{pkt}}\ \forall\ a.$

This can be verified by the following derivation:

$$\begin{aligned}
\alpha_{a'} &= \mathrm{E}\{\mathrm{E}(Y_{pkt}|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = a') - \mathrm{E}(Y_{pkt}|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = 1)\} \\
&= \mathrm{E}\{\mathrm{E}(Y_{pkt}(a')|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = a') - \mathrm{E}(Y_{pkt}(1)|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}, A_{pkt} = 1)\} \\
&\quad \text{by assumption 1.2.2} \\
&= \mathrm{E}\{\mathrm{E}(Y_{pkt}(a')|\mathbf{X_{pkt}} = \mathbf{x_{pkt}}) - \mathrm{E}(Y_{pkt}(1)|\mathbf{X_{pkt}} = \mathbf{x_{pkt}})\} \text{ by condition 2.4.1} \\
&= \mathrm{E}\{\mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1)|\mathbf{X_{pkt}} = \mathbf{x_{pkt}})\} \\
&= \mathrm{E}(Y_{pkt}(a') - Y_{pkt}(1))
\end{aligned}$$

But counterfactual condition 2.4.1 gives little intuition about the **substantive conditions** under which causal conclusions are justified from the VAM in equation 2.4.1. In contrast, Rothstein (2009) does discuss these substantive conditions, and argues that VAMs (like the one in equation 2.4.1) will produce unbiased estimates only if students are assigned to teachers on the basis of observable characteristics. Though he does not use graphs to communicate these conditions, the SWOOPs in figure 2.3.2 in section 2.3.1 illustrate the conditions identified by Rothstein (2009); i.e., students are not assigned to teachers on the basis of any unobserved characteristics (the attributes mapped to $\mathbf{U_{pk}}$ and $\mathbf{U_{pkt}}$). We showed in section 2.3.1 that the assumptions encoded by the SWOOPs in figure 2.3.2 imply that $A_{pkt} \perp\!\!\!\perp Y_{p'k't'}(a)\ \forall\ p, p', k, k', t, t', a$, which in turn implies condition 2.4.1. Thus the SWOOPs in figure 2.3.2 allow us to connect the substantive conditions in Rothstein (2009) directly to the counterfactual conditions in Reardon and Raudenbush (2009).

We can use the other SWOOPs in section 2.3 to communicate the substantive conditions and verify the counterfactual conditions that justify causal conclusions from VAMs in more complicated settings. For example, the SWOOPs in figure 2.3.4 show the substantive conditions that justify causal conclusions from a VAM estimated across multiple schools that includes school-level controls; the SWOOPs in figure 2.3.5 show the substantive conditions that justify causal conclusions from a VAM estimated across multiple schools that includes an effect for each school; the SWOOPs in figure 2.3.7 show the substantive conditions that justify causal conclusions from a VAM estimated across multiple years that includes student-level controls; and the SWOOPs in figure 2.3.8 show the substantive conditions that justify causal conclusions from a VAM estimated across multiple years that includes an effect for each student.

2.5 Conclusions

In this chapter, we demonstrate that SWOOPs allow researchers both to communicate broad assumptions about the relationships between *groups* of attributes and to establish the counterfactual independence conditions that justify causal conclusions from observational data. To our knowledge, this is the first work to combine developments from the literature on object-oriented networks (e.g. Koller and Pfeffer, 1997) with developments from the literature on plate models (Buntine, 1994) to substantially simplify the relationships in a typical DAG.

Like any graphical framework, SWOOPs have some limitations. The first is that SWOOPs do not provide a natural way to represent *non-nested* multilevel or longitudinal data; e.g., data in which students switch schools across years (note that in the SWOOPs in figure 2.3.9, for example, students are assumed to be in the same school across all T years of data). As we discuss in section 2.2.2, Directed Acyclic Probabilistic Entity Relationship (DAPER) diagrams (Heckerman et al., 2007) like figure 2.2.6 provide an equivalent representation of multilevel data, and may be more useful when data are not perfectly nested.

A second limitation of SWOOPs is that, at least in terms of the development we present in section 2.2, missing edges in SWOOPs imply the absence of an effect *for each individual* rather than the absence of an *average causal effect* at the population level. As discussed in Richardson and Robins (2013), the latter conclusion is often preferable because it can be verified by an experiment. Fortunately, section 7 of Richardson and Robins (2013) demonstrates that SWIGs can be used to represent population-level exclusion restrictions in addition to individual-level restrictions. Thus one possible extension to the development presented in this paper would be to alter the definitions in section 2.2.2 so that missing edges in stacked independent SWOOPs imply exclusion restrictions at the population rather than individual level.

A final limitation is with our use of ID variables (Rattigan and Jensen, 2010; Rattigan et al., 2011). Rattigan et al. (2011) discuss two applications of ID variables in a school setting: one in which there is a school-level *confounder*; and the other in which there is a school-level *collider*. As we discuss in appendix B, we do not believe that the second application of ID variables is internally consistent (since ID variables and the associated attributes in the same entity class must be regarded as *fixed* for the framework to make sense). We are therefore very intentional throughout our development of SWOOPs to use ID variables only in examples of the first application. That said, we offer an alternative method of discussing school-level colliders in appendix B that does not involve the use of ID variables.

To conclude, we argue that SWOOPs offer substantial improvements over existing graphical methods for causal reasoning in multivariate, multilevel, and longitudinal settings. We envision SWOOPs as having both practical and pedagogical applications; not only can they help researchers communicate and verify the assumptions that justify causal conclusions in different applied settings, but they can help instructors teach these assumptions to students and explain *why* they justify causal conclusions.

Chapter 3

Response to Intervention? Estimating the Causal Effect of Special Education Services on Student Performance

3.1 Introduction

In 1975, the U.S. Congress passed the Individuals with Disabilities Education Act (IDEA, 1975), which required public school districts to provide a free and appropriate education to students with disabilities in the least restrictive environment possible. In the subsequent three decades–which included the 2004 re-authorization of IDEA (IDEA, 2004)–there have been relatively few large-scale studies that estimate the average treatment effect (ATE) of special education services on student performance, which is surprising for two reasons. First, the federal government has made a tremendous investment in special education services, about $50 billion annually, compared to $27.3 billion for general education and $1 billion for other special services, such as Title I (Morgan et al., 2010). Given the magnitude of the investment in special education services, it is surprising that there hasn't been a greater effort to understand the impact of these services on the achievement of students with special needs. Second, since the passage of the No Child Left Behind Act (NCLB) in 2001, researchers have received unprecedented access to longitudinal datasets of annual student test scores, often linked to individual teachers, which allows them to evaluate the impact of educational interventions on student test performance (e.g. Chetty et al., 2013; Goldhaber and Theobald, 2013b; Jacob and Lefgren, 2004). Given the subsequent explosion of research that uses these longitudinal datasets, there is significant need and ample opportunity to evaluate the impact of the best-funded educational intervention in the country's history.

However, the only two large-scale, published studies that estimate the ATE of special education services on student performance come to contradictory conclusions. Hanushek et al. (2002) use a large longitudinal dataset from Texas to estimate several variants of a student fixed-effects model, and conclude that the ATE of special education services on the performance of students who receive these services is positive. On the other hand, Morgan et al. (2010) use data from the Early Childhood Longitudinal Study to estimate a number

of different covariate adjustment models, and conclude that special education services have a relatively large and negative ATE on student test performance. These disparate findings could easily be explained by a number of differences between the studies: the time period (before NCLB in Hanushek et al. (2002) and after NCLB in Morgan et al. (2010)); the grade level (later elementary and middle school in Hanushek et al. (2002) and early elementary school in Morgan et al. (2010)); the data source (Texas Public School data in Hanushek et al. (2002) and nationally representative data in Morgan et al. (2010)); and the available control variables (generally more extensive in Morgan et al. (2010) than in Hanushek et al. (2002)). However, another potential explanation for these disparate findings is that the estimation method used in at least one of these papers (a student fixed effects model in Hanushek et al. (2002) and a covariate adjustment model in Morgan et al. (2010)) results in a biased estimate of the ATE of special education services on student performance.

In section 3.2, I use longitudinal data from Washington State to replicate the methods from Hanushek et al. (2002) and Morgan et al. (2010). This replication study demonstrates that even when these methods are applied to the *exact same dataset* (and in an extension, when these methods are applied to a subset of the data that ensures that the estimates from each method are informed by the *exact same students*), the estimates from these methods lead to contradictory conclusions. Specifically, the estimates from the student fixed effects model (Hanushek et al., 2002) suggest that special education has a small, positive ATE on student performance, while the estimates from a covariate adjustment model (Morgan et al., 2010) suggest the ATE is actually large and negative (and the difference between the estimates is statistically significant). This suggests that the estimates generated by at least one, and possibly both, of these methods are biased.

In section 3.3, therefore, I use Single-World Object-Oriented Plates (SWOOPs, developed in chapter 2) to discuss the counterfactual conditions and substantive assumptions that justify each approach. Beyond illustrating the utility of SWOOPs to both communicate and verify the assumptions that justify these different estimation methods, this discussion suggests that each of these estimation methods relies on an assumption that is potentially problematic. Specifically, a student fixed effects model (Hanushek et al., 2002) assumes that students are not dynamically selected for special education services; that is, that students are not identified for special education in response to prior performance. On the other hand, a covariate adjustment model (Morgan et al., 2010) assumes that students are not selected for special education on the basis of any variable that is not observed in the dataset. I argue that these assumptions are difficult to justify, particularly since schools and districts are now directed to use a Response To Intervention (RTI) identification process that is *specifically designed* to identify students for special education services based on (a) recent performance, and (b) factors that are not observed by researchers.

There is a third analytic approach exists that relies on a different set of assumptions about student placement into special education: instrumental variables (IV) methods. IV models require an instrumental variable that does impact the probability that a student is placed in special education, but has no other relationship (direct or indirect) with student performance. In section 3.4, I argue that a funding threshold in Washington's special educa-

tion funding laws provides a plausible instrumental variable. School districts in Washington receive special education funding from the state through a placement-based system (Mahitivanichcha and Parrish, 2005) in which districts receive a fixed amount of funding for each special education student in the district. But the state's special education funding system contains a unique funding threshold; specifically, once a district has 12.7% of its students enrolled in special education, it receives no additional funding from the state even if it places additional students in special education. There have been a small number of school districts in recent years that have started the school year with more than 12.7% of their students enrolled in special education, and empirical evidence about special education funding incentives (e.g. Cullen, 2003; Dhuey and Lipscomb, 2011; Greene and Forster, 2002; Kane and Johnson, 1993) suggests that students in these districts should be less likely to be placed in special education than students in districts not beyond the state's funding threshold. I use SWOOPs in section 3.4 to illustrate the assumptions under which the position of a student's district relative to this funding threshold can be used as an instrumental variable in an IV model to estimate the ATE of special education services on student performance.

In section 3.5, I present strong evidence that students in districts that are beyond the state special education funding threshold are less likely to be placed in special education for a specific learning disability, all else equal, than students in other districts. This finding has clear policy implications, and contributes to the growing literature (e.g. Cullen, 2003; Dhuey and Lipscomb, 2011; Greene and Forster, 2002; Kane and Johnson, 1993) demonstrating that school districts respond to the incentives created by special education funding systems. When I use the funding threshold as an instrumental variable in an IV analysis, I do not find evidence that, on average, special education services have a statistically significant impact on student performance. I then offer some concluding thoughts in section 3.6.

3.2 Replication Study

I begin by describing a replication study that uses longitudinal data from Washington State to replicate the methods from Hanushek et al. (2002) and Morgan et al. (2010). The methods from Hanushek et al. (2002) have been replicated previously by Ewing (2009) and Parker (2011), but to my knowledge, this is the first replication of the methods from Morgan et al. (2010). I introduce notation in section 3.2.1, provide additional information about Hanushek et al. (2002) and Morgan et al. (2010) in section 3.2.2, describe the data I will use in section 3.2.3, and present the results of the replication study in section 3.2.4.

3.2.1 Notation

To facilitate my discussion of Hanushek et al. (2002) and Morgan et al. (2010), I introduce notation that I will use throughout the chapter. Suppose that we observe students $p = 1, \ldots, P$ over years $t = 1, \ldots, T$. Let $A_{pt} = 1$ if student p receives special education services in year t, and $A_{pt} = 0$ otherwise. Let Y_{pt} be the test score of student p at the end of year t (this will be a math score or a reading score in different applications). Now define two potential outcomes (Neyman, 1923; Rubin, 1974) for each student p in year t: $Y_{pt}(a_{pt} = 1) \equiv Y_{pt}(1)$

is the test score for student p in year t if (potentially contrary to fact) student p received special education services in year t; and $Y_{pt}(a_{pt} = 0) \equiv Y_{pt}(0)$ is the test score for student p in year t if (potentially contrary to fact) student p did not receive special education services in year t. Only one potential outcome is observed for each student and year: $Y_{pt}(1)$ if student p received special education services in year t; or $Y_{pt}(0)$ otherwise. Though neither Hanushek et al. (2002) nor Morgan et al. (2010) discusses their methods in counterfactual terms, I will assume in this section that the goal in each paper is to estimate the average treatment effect (ATE) of special education services on student test scores:

$$\mathrm{E}(Y_{pt}(1) - Y_{pt}(0)). \tag{3.2.1}$$

The ATE is the average difference between the test score students *would have* achieved if they had received special education services in year t and the outcome students *would have* achieved if they had not received special education services in year t. This notation implies the consistency and no-interference assumptions discussed in section 1.2.1 of chapter 1.

Both Hanushek et al. (2002) and Morgan et al. (2010) control for a large number of observed attributes about each student p in year t. I distinguish between two types of observed attributes:

- $\mathbf{O_p}$ = the collection of observed, time-invariant attributes for student p.
- $\mathbf{O_{pt}}$ = the collection of observed, time-variant attributes for student p in year t.

Following the conventions developed in chapter 2, I will use $\mathbf{X_{O_p}}$ and $\mathbf{X_{O_{pt}}}$ to refer to the vectors of attributes in $\mathbf{O_p}$ and $\mathbf{O_{pt}}$, respectively. Although the specific attributes in $\mathbf{X_{O_p}}$ and $\mathbf{X_{O_{pt}}}$ vary between Hanushek et al. (2002) and Morgan et al. (2010), for the purposes of this replication study, I define these vectors in terms of the variables that are available in the Washington State data described in section 3.2.3. Specifically, $\mathbf{X_{O_p}}$ includes indicators for the race and gender of student p, while $\mathbf{X_{O_{pt}}}$ includes indicators for the eligibility of student p in year t for free/reduced lunch, gifted services, homeless services, migrant student services, and English Language Learner services.

3.2.2 Prior Work

The first large-scale study to estimate the impact of special education services on student performance, Hanushek et al. (2002), uses a large longitudinal dataset from Texas in the mid-1990s. The authors provide a comprehensive descriptive analysis of special education services in the state, including distributions of disability types and transition rates into and out of special education by grade and year, but also estimate several empirical models designed to estimate the ATE in equation 3.2.1. I will focus on the subset of models that include a fixed effect for each student p (Hanushek et al. (2002) also estimate models with a school-by-year-by-grade fixed effect, and find qualitatively similar results):

$$\mathrm{E}(Y_{pt}|\mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = a_{pt}) = \alpha_0 + \boldsymbol{\alpha_1}^T \mathbf{x_{O_{pt}}} + \alpha_{TRT} a_{pt} + \alpha_p \tag{3.2.2}$$

Due to the inclusion of the student fixed effect α_p, the coefficient α_{TRT} in equation 3.2.2 is identified by students who transition into and out of the special education designation over the years of their study (note that the vector of time-invariant student characteristics $\mathbf{X_{O_p}}$ is not included in the model because it is collinear with α_p). Under linearity, homogeneity, and no interaction assumptions discussed in section 1.2 of chapter 1, α_{TRT} can be interpreted as the average difference in student test scores for students who transition into and out of special education between the years they were receiving special education services and the years they were not, all else equal. In section 3.3, I will discuss the counterfactual and substantive conditions under which α_{TRT} is equal to the ATE in equation 3.2.1.

Hanushek et al. (2002) estimate the model in equation 3.2.2 for different disability types (e.g., learning disabled and emotionally disturbed) and transition types (entry and exit) and generally find modest, positive, and statistically-significant special education effects. For example, in the specific model that we will replicate in section 3.2.4, where we focus exclusively on students receiving special education services for a specific learning disability, Hanushek et al. (2002) estimate that the ATE of special education services on student math performance is .04 standard deviations of student performance ($T = 4.17$).

More recently, Morgan et al. (2010) use data from 2002 and 2004 waves of the Early Childhood Longitudinal Study and compare the performance of students placed into special education with the performance of "similar students" not receiving special education services. The authors identify this comparison group using a number of different covariate-adjustment models (e.g., OLS linear regression and propensity score models), all of which control for 35 different observed covariates for each student. For ease of exposition (and comparison with the methods in Hanushek et al. (2002)), I will focus on the linear regression models in Morgan et al. (2010) (they also estimate propensity score models (Rosenbaum and Rubin, 1983) and find qualitatively similar results):

$$\mathrm{E}(Y_{pt}|\mathbf{X_{O_p}} = \mathbf{x_{O_p}}, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = a_{pt}) = \beta_0 + \boldsymbol{\beta_1}^T\mathbf{x_{O_p}} + \boldsymbol{\beta_2}^T\mathbf{x_{O_{pt}}} + \beta_{TRT}a_{pt} \quad (3.2.3)$$

The vector $\mathbf{X_{O_{pt}}}$ in Morgan et al. (2010) (as in our replication in section 3.2.4) includes a measure of baseline academic performance, which will be important in our subsequent discussion. Under the same linearity, homogeneity, and no interaction assumptions discussed in section 1.2 of chapter 1, β_{TRT} can be interpreted as the average difference in performance between students who are receiving special education services and "similar" students who are not receiving special education services. Thus the key conceptual difference between α_{TRT} from Hanushek et al. (2002) and β_{TRT} from Morgan et al. (2010) is that α_{TRT} is identified by variation *within students*, while β_{TRT} is identified by variation *across students*.

Using this approach, Morgan et al. (2010) estimate special education effects that are relatively large and negative. For example, their reported estimates of β_{TRT} from the model in equation 3.2.3 are -6.622 in reading and -3.350 in math ($p < .001$ for both). Given the reported standard deviations of student test scores in their summary statistics, these estimates suggest that the ATE of special education services on student reading performance is -.32 standard deviations of student performance in reading and -.17 standard deviation of

student performance in math. The estimates from the propensity score models in Morgan et al. (2010) are somewhat smaller and not consistently statistically significant, but the overall conclusion is that special education may actually have a negative impact on student academic performance.

3.2.3 Data

As I discuss in section 3.1, the differences between the estimates from Hanushek et al. (2002) and Morgan et al. (2010) could be due to a number of important differences between the studies *other than* their analytic approaches. Specifically, the data in these studies come from different time periods, grade levels, geographic areas, and include different control variables. To isolate differences that are due to the analytic approaches employed in each study, then, it is important to replicate these methods using the same dataset.

The student-level data used in this replication study and throughout this paper were provided by the Washington State Office of the Superintendent of Public Instruction (OSPI). The state's CEDARS data system, introduced for the 2009-10 school year, contains detailed demographic information about each public school student in the state, linkable to a special education file that contains the date of entry, disability type, and (if applicable) date of exit of every student in the state who has received special education services. These data are also linkable to student test score data from the Measures of Student Progress (MSP) exam, an annual exam in math and reading given to every student in grades 3-8 in the state. I standardize these test scores within grade, year, and school district so that each test score Y_{pt} can be interpreted as the performance of student p in year t *relative to other students in the same year, grade, and district*. I discuss the consequences of this parameterization of test performance throughout the paper.

Throughout this chapter, I focus exclusively on one disability type within special education: specific learning disabilities (although I use other disability types to perform a robustness check described in section 3.5.3). I focus on this disability type for two primary reasons. First, specific learning disabilities represent the largest category of disability type within special education, and unlike the second-largest category of special education (speech or language impediments), special education services for a specific learning disability are *specifically intended* to improve student performance in school. Second, the identification strategy described in section 3.4 relies on administrators having some flexibility in which and how many students they assign to special education. The rules governing placement in special education for a specific learning disability are generally "more subjective" (Goldstein, 2003) than the rules governing placement into other categories of special education, so it is plausible that the probability that a student is placed in special education for a specific learning disability can be influenced by outside factors (such as the funding threshold described in section 3.4). Throughout, I focus exclusively on students in grades 4-8 (since students in these grade levels have prior-year test scores that are highly predictive of special education placement for a specific learning disability), and use data from four school years (2009-10 through 2012-13).

For the replication study, I create a dataset that is closely analogous to the longitudinal dataset in Hanushek et al. (2002), but can also be used to replicate the methods in Morgan et al. (2010). In this dataset, I set $A_{pt} = 1$ for student p in year t if the student received special education services for a specific learning disability for at least half of year t. This analytic dataset contains 1,684,849 student/year observations (of which $A_{pt} = 1$ for 107,460 student/year observations) and 616,954 unique students (of whom 45,203 students receive special education for a specific learning disability during one of the years of the study). Importantly, 13,287 of these students transition between special education for a specific learning disability and general education during the years of data. These students identify the coefficient α_{TRT} from the student fixed effects model in equation 3.2.2.

3.2.4 Replication Results

I use the dataset described in section 3.2.3 to estimate equations 3.2.2 and 3.2.3 and replicate, to the extent possible with the available data, the methods and empirical specifications of Hanushek et al. (2002) and Morgan et al. (2010), respectively. Table 3.2.1 displays the coefficient estimates from these models. The first two columns display estimates from models in which the dependent variable Y_{pt} is student performance in reading, and the second two columns display analogous estimates for math. The columns labeled "SFE" contain estimates from the student fixed-effects model (equation 3.2.2, replicating Hanushek et al. (2002)), while the columns labeled "COV" contain estimates from the covariate adjustment model (equation 3.2.3, replicating Morgan et al. (2010)).

The row "LD services" in table 3.2.1 contains the estimates of the treatment effects α_{TRT} (from the student fixed effects model) and β_{TRT} (from the covariate adjustment model). These estimates are extremely consistent with the findings from Hanushek et al. (2002) and Morgan et al. (2010). Specifically, the estimates from the student fixed effects model suggest that special education for a specific learning disability has a modest positive (in reading) or statistically insignificant (in math) effect on student performance, which broadly parallels the findings from Hanushek et al. (2002). On the other hand, the estimates from the covariate adjustment model suggest that special education for a specific learning disability has a large negative effect on both math and reading student performance, which broadly parallels the findings from Morgan et al. (2010). Not only are the direction of these estimates consistent with the estimates from Hanushek et al. (2002) and Morgan et al. (2010), but the magnitudes are strikingly similar to the estimates discussed in section 3.2.2.

The primary conclusion from this replication study is that the differences between the estimates from Hanushek et al. (2002) and Morgan et al. (2010) do *not* appear to be due to the different time periods, grade levels, geographic areas, and control variables in these papers. Instead, the estimation methods in these papers produce contradictory conclusions about the effects of special education even when they are applied to the exact same dataset. In section 3.3, I discuss the counterfactual conditions and substantive assumptions that justify each approach.

	Reading		Math	
	SFE	**COV**	**SFE**	**COV**
Intervention (A_{pt})				
LD services	0.046***	-0.205***	0.005	-0.159***
	(0.011)	(0.002)	(0.009)	(0.002)
Time-variant covariates ($\mathbf{O_{pt}}$)				
FRL eligibility	0.001	-0.077***	0.006	-0.069***
	(0.005)	(0.001)	(0.004)	(0.001)
Gifted services	-0.046***	0.167***	-0.063***	0.231***
	(0.006)	(0.002)	(0.005)	(0.002)
ELL services	0.061***	-0.141***	-0.015*	-0.029***
	(0.008)	(0.003)	(0.007)	(0.002)
Migrant services	-0.012	0.009*	0.013	0.043***
	(0.013)	(0.004)	(0.011)	(0.004)
Homeless services	-0.003	-0.048***	-0.015	-0.057***
	(0.010)	(0.004)	(0.009)	(0.004)
Time-invariant covariates ($\mathbf{O_p}$)				
Female		0.138***		-0.027***
		(0.001)		(0.001)
American Indian		-0.089***		-0.059***
		(0.005)		(0.005)
Asian / Pacific Islander		0.029***		0.124***
		(0.002)		(0.002)
Black		-0.023***		-0.076***
		(0.003)		(0.003)
Hispanic		-0.008***		-0.015***
		(0.002)		(0.002)
Controls for prior test scores	No	Yes	No	Yes
Student fixed effect	Yes	No	Yes	No

*$p < .05$; **$p < .01$; ***$p < .001$

Table 3.2.1: Estimated coefficients from replication study on full dataset

3.3 Discussion

Much like the discussion in section 1.3 of chapter 1, I now identify and connect the counterfactual conditions (section 3.3.1) and substantive assumptions (section 3.3.2) that justify causal conclusions from the models in Hanushek et al. (2002) and Morgan et al. (2010). This section builds on the discussion in chapter 1 in three ways. First, instead of the stylized example developed in chapter 1, this discussion concerns a real-life example of Lord's Paradox (in which two reasonable approaches to estimating the same effect come to very different conclusions). As such, this example contains considerably more potential confounders than the one "observed" and one "unobserved" confounder in chapter 1. Because of this, I use SWOOPs (developed in chapter 2) rather than SWITs to communicate and verify the assumptions that justify causal conclusions from these models. This application of SWOOPs is intended to illustrate the utility of this new graphical framework in a real-life application.

3.3.1 Counterfactual Conditions

I begin by discussing the student fixed effects model in equation 3.2.2 (Hanushek et al., 2002). Building on the discussion of "ID variables" in chapter 2, I first define a student-level ID variable $I_p = p$ for each student $p = 1, \ldots, P$. With this notation, the coefficient α_{TRT} in equation 3.2.2 can be derived directly from equation 3.2.2:

$$\alpha_{TRT} = \mathrm{E}\left\{\mathrm{E}(Y_{pt}|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 0)\right\}.$$

As discussed in section 3.2.2, this estimand can be interpreted as the expected difference in student test scores for students who transition into and out of special education between the years they were receiving special education services and the years they were not, all else equal. This estimand equals the ATE $\mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$ under the following counterfactual condition (as discussed in section 2.3.3 of chapter 2, the notation $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$, where $\overline{\mathbf{a}}_{\mathbf{pt'}} = \{a_{p1}, \ldots, a_{pt'}\}$ for $t' = 1, \ldots, T$, reflects the reality that a student's performance in year t may depend on whether they've received special education services in previous years as well):

Condition 3.3.1 *(Justifying Hanushek et al. (2002))* $A_{pt} \perp\!\!\!\perp Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})|I_p, \mathbf{X_{O_{pt}}}\ \forall\ \overline{\mathbf{a}}_{\mathbf{pt}}, p, t.$

In the language of chapter 1, this condition implies that assignment to special education must be conditionally ignorable given the ID variable and observed, time-variant covariates. The following derivation demonstrates why condition 3.3.1 implies that the quantity estimated by Hanushek et al. (2002) (α_{TRT}) equals the ATE $\mathrm{E}(Y_2(1) - Y_2(0))$:

$$\begin{aligned}
\alpha_{TRT} &= \mathrm{E}\left\{\mathrm{E}(Y_{pt}|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 0)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1)|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}(0)|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}, A_{pt} = 0)\right\} \\
&\quad \text{by assumption 1.2.2} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1)|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}}) - \mathrm{E}(Y_{pt}(0)|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}})\right\} \text{ by condition 3.3.1} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1) - Y_{pt}(0)|I_p = p, \mathbf{X_{O_{pt}}} = \mathbf{x_{O_{pt}}})\right\} \\
&= \mathrm{E}(Y_{pt}(1) - Y_{pt}(0))
\end{aligned}$$

I now turn to the covariate adjustment model in equation 3.2.3 (Morgan et al., 2010). I first define the vector $\mathbf{X_{pt}} = \{Y_{p(t-1)}, \mathbf{X_{O_p}}, \mathbf{X_{O_{pt}}}\}$; i.e., $\mathbf{X_{pt}}$ is the set of all observed covariates for student p and in year t (including the student's lagged test score). With this notation, the coefficient β_{TRT} in equation 3.2.3 is equal to:

$$\beta_{TRT} = \mathrm{E}\left\{\mathrm{E}(Y_{pt}|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 0)\right\}. \tag{3.3.1}$$

As discussed in section 3.2.2, this estimand can be interpreted as the expected difference in performance between students who are receiving special education services and "similar" students who are not receiving special education services. This estimand equals the ATE $\mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$ under the following counterfactual condition:

Condition 3.3.2 *(Justifying Morgan et al. (2010))* $A_{pt} \perp\!\!\!\perp Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})|\mathbf{X_{pt}} \; \forall \; \overline{\mathbf{a}}_{\mathbf{pt}}, p, t.$

This condition implies that assignment to special education must be conditionally ignorable given all observed student covariates (including lagged performance). The following derivation demonstrates why condition 3.3.2 implies that the quantity estimated by Morgan et al. (2010) (β_{TRT} in equation 3.3.1) equals the ATE $\mathrm{E}(Y_2(1) - Y_2(0))$:

$$\begin{aligned}
\beta_{TRT} &= \mathrm{E}\left\{\mathrm{E}(Y_{pt}|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 0)\right\} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1)|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 1) - \mathrm{E}(Y_{pt}(0)|\mathbf{X_{pt}} = \mathbf{x_{pt}}, A_{pt} = 0)\right\} \text{ by consistency} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1)|\mathbf{X_{pt}} = \mathbf{x_{pt}}) - \mathrm{E}(Y_{pt}(0)|\mathbf{X_{pt}} = \mathbf{x_{pt}})\right\} \text{ by condition 3.3.2} \\
&= \mathrm{E}\left\{\mathrm{E}(Y_{pt}(1) - Y_{pt}(0)|\mathbf{X_{pt}} = \mathbf{x_{pt}})\right\} \\
&= \mathrm{E}(Y_{pt}(1) - Y_{pt}(0))
\end{aligned}$$

The key question, then, is under what substantive assumptions do conditions 3.3.1 and 3.3.2 hold? In section 3.3.2, I use SWOOPs to illustrate these assumptions and connect them to conditions 3.3.1 and 3.3.2.

3.3.2 Substantive Assumptions

As in section 2.3 of chapter 2, I define two sets of *unobserved* covariates that will facilitate the discussion of the substantive assumptions justifying the methods in Hanushek et al. (2002) and Morgan et al. (2010).

- $\mathbf{U_p}$ = the collection of unobserved, time-invariant characteristics for student p (e.g., IQ).
- $\mathbf{U_{pt}}$ = the collection of unobserved, time-variant characteristics for student p in year t (e.g., parental involvement).

Figure 3.3.1 contains stacked SWOOPs that illustrate the assumed relationships (black edges) and potential relationships (blue edges) between the four collections of student covariates, the intervention, and the outcome in these studies.

Recall from chapter 2 that the SWOOPs in figure 3.3.1 show relationships between *collections* of attributes, the intervention (special education services), and the outcome (test

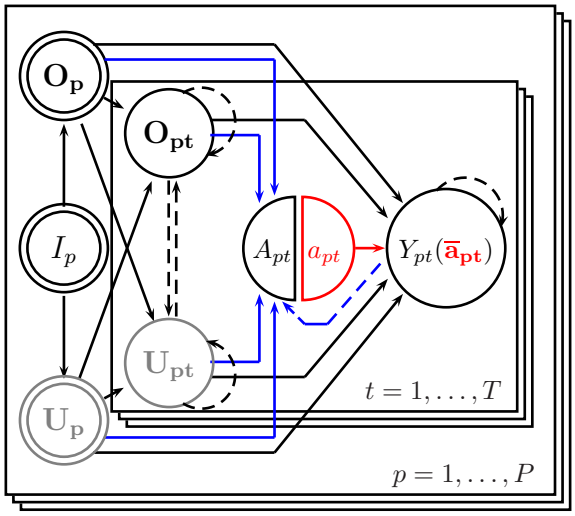


Figure 3.3.1: Stacked SWOOPs for special education placement

performance). For example, the solid edges between $Y_{pt}(a_{pt})$ and $\mathbf{O_p}$, $\mathbf{O_{pt}}$, $\mathbf{U_p}$, and $\mathbf{U_{pt}}$ represents the assumption that *at least one* attribute in each of these collections of attributes (e.g., gender, ELL services, IQ, and parental involvement, respectively) influences test performance. Further recall that the dashed edges in the SWOOPs in figure 3.3.1 represent relationships between groups of attributes in year $t-1$ and groups of attributes in year t. For example, the dashed edge from U_{pt} to O_{pt} represents the assumption that at least one unobserved, time-variant attribute in year $t-1$ (e.g., English proficiency) influences at least one observed, time-variant attribute in year t (e.g., ELL classification). Finally, recall that the presence of the student ID variable I_p in the SWOOPs in figure 3.3.1 means that the objects O_p and U_p are both deterministic objects (since all time-invariant student attributes are deterministic functions of the student ID variable).

As in chapter 2, I will focus in this section on the four blue edges in figure 3.3.1 between objects representing different collections of attributes and the intervention node A_{pt}. The *absence* of these edges illustrate the substantive assumptions we will discuss in this section about how students are identified for special education services. For example, under the assumption that students are not identified for special education on the basis of any unobserved, time-variant attributes, the edge between $\mathbf{U_{pt}}$ and A_{pt} is missing. Likewise, the dashed edge from the outcome object $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ to the intervention object A_{pt} is missing if and only if students are not selected for special education on the basis of their test performance in year $t-1$.

I now use these stacked SWOOPs to illustrate the substantive assumptions under which conditions 3.3.1 and 3.3.2 hold (and thus under which the methods from Hanushek et al. (2002) and Morgan et al. (2010) are justified). Figure 3.3.2 illustrates the substantive assumptions that imply condition 3.3.1 (and thus justify the methods from Hanushek et al. (2002)). The missing edge between U_{pt} and A_{pt} represents the assumption that students are not selected for special education on the basis of any unobserved, time-variant attributes.

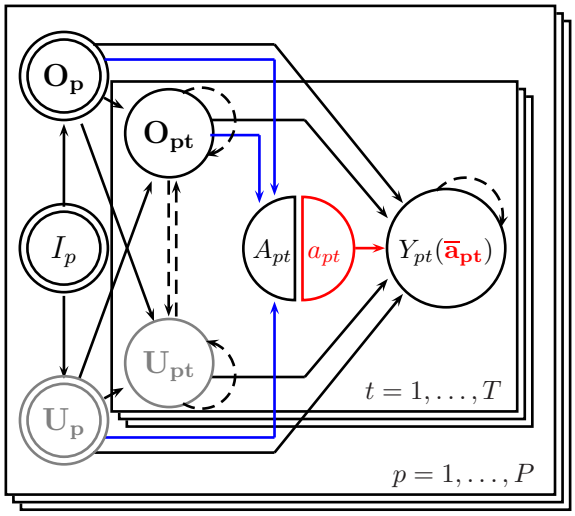


Figure 3.3.2: Substantive assumptions justifying student fixed effects model (Hanushek et al., 2002)

Likewise, the missing dashed edge from the outcome object $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ to the intervention object A_{pt} represents the assumptions that students are not selected for special education on the basis of their test performance in year $t-1$.

In the SWOOPs in figure 3.3.2, every path between A_{pt} and $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ contains the non-collider $\mathbf{O_{pt}}$, $\mathbf{O_p}$, or $\mathbf{U_p}$. By the definitions outlined in chapter 2, A_{pt} and $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ are D-separated given the set $S = \{\mathbf{O_{pt}}, I_p\}$ in figure 3.3.2. By the theorems in chapter 2, this implies that for any $\overline{a}_{pt}$, $A_{pt} \perp\!\!\!\perp Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}}) \mid \mathbf{X}_{\mathbf{O_{pt}}}, I_p$ in any distribution P that is Markov with respect to the underlying SWIT $\mathcal{G}$ (as in chapter 2, I will omit this qualification from now on and just assume that the Markov property holds in any underlying SWIT). This is precisely condition 3.3.1 from section 3.3.2. I have therefore demonstrated that the substantive assumptions illustrated in figure 3.3.2 imply condition 3.3.1, which in turn implies that the coefficient α_{TRT} in equation 3.2.2 is equal to the ATE. Thus the methods in Hanushek et al. (2002) are justified under the the substantive assumptions illustrated in figure 3.3.2.

On the other hand, figure 3.3.3 illustrates the substantive assumptions that imply condition 3.3.2 (and thus justify the methods from Morgan et al. (2010)). The missing edges between U_{pt} and A_{pt} and between U_p and A_{pt} represent the assumption that students are not selected for special education on the basis of any unobserved covariates (time-invariant or time-variant).

In the SWOOPs in figure 3.3.3, every path between A_{pt} and $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ contains the non-collider $\mathbf{O_{pt}}$, $\mathbf{O_p}$, or $Y_{p(t-1)}(\overline{\mathbf{a}}_{\mathbf{p(t-1)}})$. Thus A_{pt} and $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ are D-separated given the set $\{\mathbf{O_{pt}}, \mathbf{O_p}, Y_{p(t-1)}(\overline{\mathbf{a}}_{\mathbf{p(t-1)}})\}$, which in turn implies that for any a_{pt}, $A_{pt} \perp\!\!\!\perp Y_{pt}(a_{pt}) \mid \mathbf{X_{pt}} = \{\mathbf{X}_{\mathbf{O_{pt}}}, \mathbf{X}_{\mathbf{O_p}}, Y_{p(t-1)}(\overline{\mathbf{a}}_{\mathbf{p(t-1)}})\}$. This is precisely condition 3.3.2 from section 3.3.2, so the substantive assumptions illustrated in figure 3.3.3 imply condition 3.3.2 and the coefficient β_{TRT} in equation 3.2.3 is equal to the ATE. Thus the methods in Morgan et al. (2010) are justified

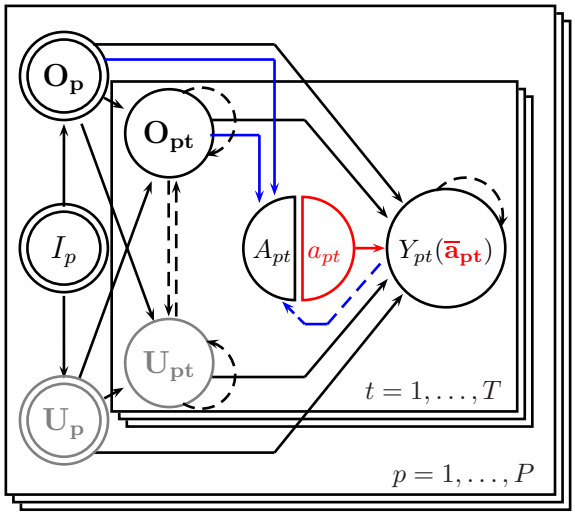


Figure 3.3.3: Substantive assumptions justifying covariate adjustment model (Morgan et al., 2010)

under the the substantive assumptions illustrated in figure 3.3.3.

As we discuss in chapter 1, it is not possible to test these substantive assumptions using the observed data. However, in section 3.3.3, I briefly outline why there are good reasons to doubt that *either* of these sets of substantive assumptions hold in this application. This motivates the IV approach I describe in section 3.4.

3.3.3 Threats to Causal Conclusions

Federal regulations for identifying students for special education services (IDEA 1975, 2004) call the substantive assumptions discussed in section 3.3.3 into question. For example, schools and districts are now directed to use a "Response To Intervention" (RTI) identification process that identifies a student for special education services based on his or her response to scientific, research-based intervention. In other words, the RTI placement system is *specifically designed* to identify students for special education services in response to (a) recent performance, and (b) variables that are not observed to researchers. Thus it is very difficult to argue that the solid edges between A_{pt} and $\mathbf{U_p}$ and $\mathbf{U_{pt}}$ or the dashed edge between A_{pt} and $Y_{pt}(a_{pt})$ in figure 3.3.1 should be missing.

Perhaps more disconcertingly, the biases that result from violations of these substantive assumptions are likely to be in the direction of the findings of each paper (and in the direction of the estimates from the replication study). For Hanushek et al. (2002), consider figure 3.3.4, which shows the performance of two hypothetical students over time. If Student B is more likely than Student A to be placed in special education in year 3 *because* he or she performed unusually poorly in year 2, then in a student fixed effect model, Student B's regression to the mean in year 3 gets attributed to the effect of special education services. The resulting bias is therefore likely to be positive in the student fixed effects model in equation

3.2.2, and the estimated effect of special education in Hanushek et al. (2002) may be biased upwards.

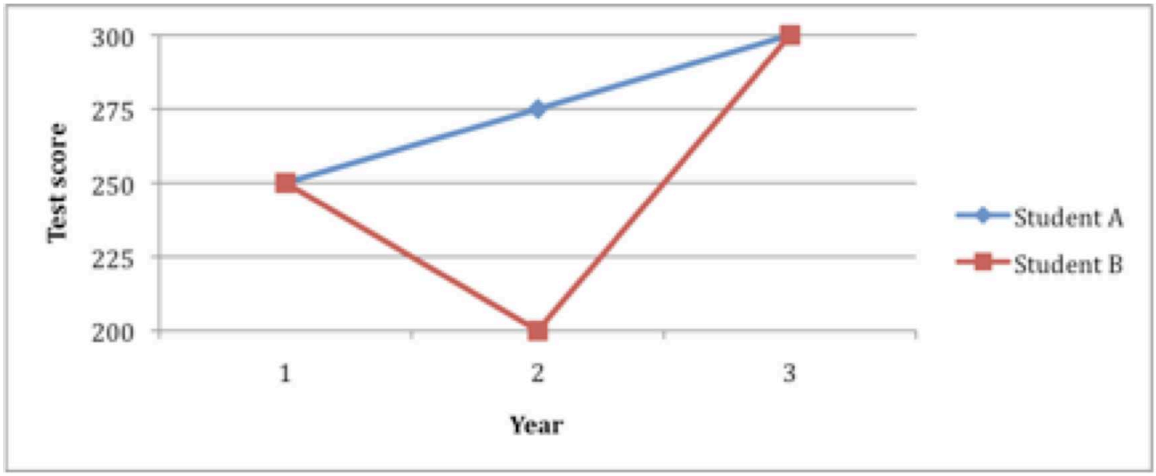


Figure 3.3.4: Conceptual figure for student fixed effects model (Hanushek et al., 2002)

On the other hand, the most likely source of bias in the covariate adjustment model from Morgan et al. (2010) is omitted variable bias. A few particularly likely confounders include student IQ (part of $\mathbf{U_p}$) and the student's parental involvement and motivation in school (part of $\mathbf{U_{pt}}$). Each of these confounders is likely to be *negatively* correlated with the probability of special education placement, but *positively* correlated with student test performance. Thus the resulting "omitted variable bias" is likely to be negative in the covariate adjustment model in equation 3.2.3, and the estimated effect of special education in Morgan et al. (2010) may be biased downwards. I would argue that the assumptions that justify a covariate adjustment model (Morgan et al., 2010) seem even less plausible than the assumptions that justify a student-fixed effects model (Hanushek et al., 2002).

3.4 A New Approach

Given the concerns outlined in section 3.3.3, it is difficult to know what to make of the disparate findings in the existing literature on special education program effects. Ideally, I would not need to make either set of substantive assumptions described in section 3.3.2 in estimating the effect of special education services on student performance. A third method–instrumental variables (IV) analysis (e.g. Angrist et al., 1996)–does not rely on either set of assumptions. Instead, IV methods require an instrumental variable that is correlated with a student's probability of special education placement, but has no other correlation (direct or indirect) with student test performance. In section 3.4.1, I argue that a funding threshold in Washington State's special education funding system provides a plausible instrumental variable. Following the same structure as section 3.3, I then outline the counterfactual

conditions that justify the IV approach in section 3.4.2, and use SWOOPs in section 3.4.3 to discuss the substantive assumptions under which these counterfactual conditions hold.

3.4.1 Background: Special Education Funding

A growing literature investigates whether school districts respond to the incentives created by different special education funding systems. Mahitivanichcha and Parrish (2005) categorize state special education funding systems into two predominant categories: "census-based" systems in which districts receive a fixed amount of money for special education services regardless of the number of special education students in the district; and "placement-based systems" in which districts receive a fixed amount of money per special education student. Some policymakers have worried that placement-based systems may incentivize districts to identify more students for special education services than they would in a census-based system (Goldstein, 2003).

A number of law changes have allowed researchers to put this theory to the test. Kane and Johnson (1993) and Dhuey and Lipscomb (2011) investigate special education placement rates in states that change from a placement-based system to a census-based system (in Vermont and California, respectively), and find that special education enrollments dropped under the census-based system. Cullen (2003) investigates the transition in Texas from a census-based system to a placement-based system and finds higher enrollments in the placement-based system. Greene and Forster (2002) also find that states with placement-based systems have higher special education rates, all else equal, than states with census-based systems. Together, this literature paints a consistent and intuitive picture: districts tend to respond to the incentives created by state special education funding systems, and tend to place more students in special education when they receive additional money from the state to provide services for these students.

Washington State offers an interesting hybrid of these special education funding systems. School districts in Washington state receive a Basic Education Allocation (typically around $5000) from the state for each student in the district, and receive additional special education funding through a placement-based system in which a fixed amount (equaling 0.9309 times the districts BEA) is provided for each student in the district that is enrolled in special education. However, according to state school funding laws in place for over a decade, this additional special education allocation is capped at 12.7% of the district's students. That is, if a district already has more than 12.7% of its students enrolled in special education, it does not receive any additional special education funds from the state regardless of how many additional students in the district get placed in special education.

Given the empirical evidence that school districts respond to the incentives created by special education funding systems, it is reasonable to expect the funding threshold to create an incentive for school districts to keep special education enrollment rates below 12.7%. However, there have been a small number of school districts in recent years that have started the school year with more than 12.7% of their students enrolled in special education, meaning that these districts did not receive any additional funding for students they placed in special

education over the course of the school year. The literature on special education funding incentives suggests that these districts should be less likely to place students in special education in these years. In sections 3.4.2 and 3.4.3, I discuss the counterfactual conditions and substantive assumptions that justify using the position of a student's district relative to this funding threshold as an instrumental variable to estimate the ATE of special education services on student performance.

3.4.2 Counterfactual Conditions

For this discussion of the counterfactual conditions that justify an IV analysis, I introduce one new variable and (following Imbens and Angrist (1994)) make one notational change from section 3.2.1. Specifically, I define the *instrumental variable* Z_{pt}, and re-define the special education indicator A_{pt} as a potential outcome that depends on the value z_{pt} of Z_{pt}.

- $Z_{pt} = 1$ if student p attends a district in year t that is beyond that state's funding threshold at the end of September (and $Z_{pt} = 0$ otherwise).
- $A_{pt}(z_{pt} = 1) \equiv A_{pt}(1) = 1$ if student p is placed in special education in year t if (potentially contrary to fact) $Z_{pt} = 1$ (and $A_{pt}(1) = 0$ otherwise).
- $A_{pt}(z_{pt} = 0) \equiv A_{pt}(0) = 1$ if student p is placed in special education in year t if (potentially contrary to fact) $Z_{pt} = 0$ (and $A_{pt}(0) = 0$ otherwise).

As before, the notation $A_{pt}(z_{pt})$ implies consistency and no-interference assumptions discussed in section 1.2.1 of chapter 1. As in section 3.3.1, define the vector $\mathbf{X_{pt}} = \{Y_{p(t-1)}, \mathbf{X_{O_p}}, \mathbf{X_{O_{pt}}}\}$. With these definitions, the estimand in an IV analysis can be expressed as:

$$\beta^{IV} = \frac{\mathrm{E}\{\mathrm{E}(Y_{pt}|Z_{pt} = 1, \mathbf{X_{pt}} = \mathbf{x_{pt}}) - \mathrm{E}(Y_{pt}|Z_{pt} = 0, \mathbf{X_{pt}} = \mathbf{x_{pt}})\}}{\mathrm{E}\{\mathrm{E}(A_{pt}|Z_{pt} = 1, \mathbf{X_{pt}} = \mathbf{x_{pt}}) - \mathrm{E}(A_{pt}|Z_{pt} = 0, \mathbf{X_{pt}} = \mathbf{x_{pt}})\}}.$$

Since A_{pt} is binary in this application, the denominator of β^{IV} is the expected difference between the probability that $A_{pt} = 1$ when $Z_{pt} = 1$ and the probability that $A_{pt} = 1$ when $Z_{pt} = 0$, holding values of $\mathbf{X_{pt}}$ constant. To ensure that this denominator does not equal zero, IV analysis requires that the value z_{pt} of Z_{pt} be predictive of A_{pt}, all else equal:

Condition 3.4.1 *$E(A_{pt}|Z_{pt} = z_{pt}, \mathbf{X_{pt}} = \mathbf{x_{pt}})$ is a non-trivial function of z_{pt}.*

Condition 3.4.1 is *testable* from observed data, and I will show in section 3.5.3 that the instrumental variable in this application is, in fact, quite predictive of whether an individual student is placed into special education for a specific learning disability.

On the other hand, IV analyses depend on two other counterfactual conditions that are not directly testable from the observed data. First, given the notation above, the outcome Y_{pt} could be a function of both the value a_{pt} of A_{pt} and the value z_{pt} of Z_{pt}. The *exclusion restriction* requires that the outcome be a function *only* of a_{pt}.

Condition 3.4.2 *(Exclusion restriction) $Y_{pt}(a_{pt}, z_{pt}) = Y(a_{pt}) \; \forall \; p, t$*

In other words, the value of the instrumental variable cannot have a *direct* effect on the outcome. In this application, condition 3.4.2 requires that a student's test performance is not a function of whether or not the student attends a district that is beyond the state's special education funding threshold.

While condition 3.4.2 is relatively easy to justify in this application, an IV analysis also requires that there is no confounding between the instrumental variable and the potential outcomes for the treatment and outcome:

Condition 3.4.3 $Z_{pt} \perp\!\!\!\perp Y_{pt}(a_{pt}), A_{pt}(z_{pt}) \mid \mathbf{X_{pt}}$ *for* $a_{pt}, z_{pt} \in \{0, 1\}$ *and* $\forall\, p, t$.

Essentially, condition 3.4.3 requires that the instrumental variable not have any *indirect* correlation with the treatment or outcome, controlling for observable attributes. This takes more work to justify in this application, so I will focus primarily on the substantive conditions that justify condition 3.4.3 in section 3.4.3.

Finally, an IV analysis requires at least one other assumption. For the IV estimand β^{IV} to equal the ATE $\mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$, I need to assume that the treatment effect of special education is constant across students and years:

Condition 3.4.4 $Y_{pt}(1) - Y_{pt}(0) = E(Y_{pt}(1) - Y_{pt}(0)) \;\forall\, p, t$.

Conditions 3.4.1-3.4.4 imply that $\beta^{IV} = \mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$. I first focus on the numerator of β^{IV}:

$$
\begin{aligned}
& \mathrm{E}\{\mathrm{E}(Y_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(Y_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \; & \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1) + (1-A_{pt}(1))Y_{pt}(0)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& - \mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1) + (1-A_{pt}(0))Y_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& \text{by condition 3.4.2} \\
= \; & \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& + \mathrm{E}\{(1-A_{pt}(1))Y_{pt}(0)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& - \mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& + \mathrm{E}\{(1-A_{pt}(0))Y_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
= \; & \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& + \mathrm{E}\{(1-A_{pt}(1))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& - \mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& + \mathrm{E}\{(1-A_{pt}(0))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \text{ by condition 3.4.3} \\
= \; & \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1) + (1-A_{pt}(1))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& - \mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1) + (1-A_{pt}(0))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
= \; & \mathrm{E}\{A_{pt}(1)Y_{pt}(1) + (1-A_{pt}(1))Y_{pt}(0)\} \\
& - \{\mathrm{E}\{A_{pt}(0)Y_{pt}(1) + (1-A_{pt}(0))Y_{pt}(0)\} \\
= \; & \mathrm{E}\{(A_{pt}(1)Y_{pt}(1) - A_{pt}(1)Y_{pt}(0) - A_{pt}(0)Y_{pt}(1) + A_{pt}(0)Y_{pt}(0)\} \\
= \; & \mathrm{E}\{(A_{pt}(1) - A_{pt}(0))(Y_{pt}(1) - Y_{pt}(0))\} \\
= \; & \mathrm{E}(A_{pt}(1) - A_{pt}(0))\mathrm{E}(Y_{pt}(1) - Y_{pt}(0)) \text{ by condition 3.4.4.}
\end{aligned}
$$

I can then derive a similar expression for the denominator of β^{IV}:

$$
\begin{aligned}
& \mathrm{E}\{\mathrm{E}(A_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(A_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \; & \mathrm{E}\{\mathrm{E}(A_{pt}(1)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(A_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \; & \mathrm{E}\{\mathrm{E}(A_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(A_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \text{ by condition 3.4.3} \\
= \; & \mathrm{E}\{\mathrm{E}(A_{pt}(1) - A_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \; & \mathrm{E}(A_{pt}(1) - A_{pt}(0)) \neq 0 \text{ by condition 3.4.1.}
\end{aligned}
$$

Finally, I can put the numerator and denominator of β^{IV} together:

$$
\begin{aligned}
\beta^{IV} &= \frac{\mathrm{E}\{\mathrm{E}(Y_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(Y_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\}}{\mathrm{E}\{\mathrm{E}(A_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}) - \mathrm{E}(A_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\}} \\
&= \frac{\mathrm{E}(A_{pt}(1) - A_{pt}(0))\mathrm{E}(Y_{pt}(1) - Y_{pt}(0))}{\mathrm{E}(A_{pt}(1) - A_{pt}(0))} \\
&= \mathrm{E}(Y_{pt}(1) - Y_{pt}(0)).
\end{aligned}
$$

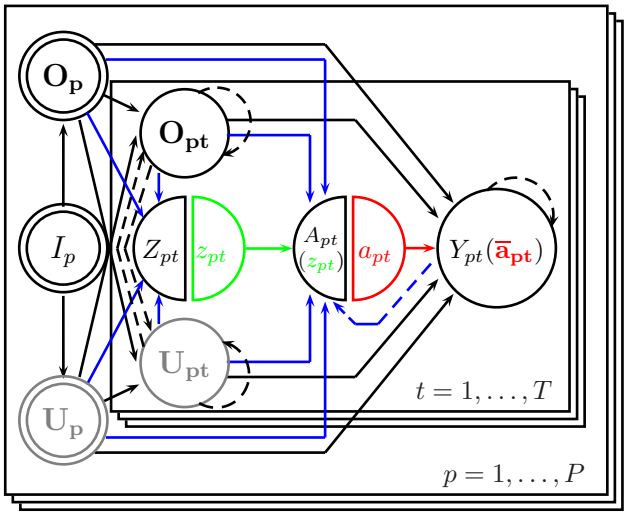


Figure 3.4.1: Stacked SWOOPs with IV variable

Though necessary to prove that $\beta^{IV} = \mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$, I would argue that condition 3.4.4 is not very plausible in this application. Specifically, it is difficult to argue that special education services will have the exact same effect on all students. So, in appendix C, I introduce a condition called monotonicity (Imbens and Angrist, 1994) to demonstrate that, under this condition (rather than condition 3.4.4) and conditions 3.4.1-3.4.3, the IV estimand β^{IV} equals the Local Average Treatment Effect $E(Y_{pt}(1) - Y_{pt}(0)|A_{pt}(1) \neq A_{pt}(0))$. The bottom line, though, is that conditions 3.4.1-3.4.3 are essential regardless of the causal estimand of interest. In section 3.4.3, therefore, I use SWOOPs to discuss the substantive assumptions under which these three conditions hold.

3.4.3 Substantive Assumptions

To facilitate this discussion of the substantive assumptions that justify an IV approach, I modify the SWOOPs from section 3.3 in three ways. First, I introduce a second split "treatment object" for the instrumental variable Z_{pt}. Second, I change the special education node A_{pt} to be a potential outcome that depends on the value z_{pt} of Z_{pt}. Finally, I add additional blue edges that represent potential relationships between groups of attributes and the instrumental variable Z_{pt}. Figure 3.4.1 contains the resulting SWOOPs.

The relationships displayed in the SWOOPs in figure 3.4.1 already illustrate two of the three conditions discussed in section 3.4.2. Specifically, the edge between z_{pt} and $A_{pt}(z_{pt})$ indicates that the value of the instrumental variable Z_{pt} influences the probability that a student is placed in special education (i.e., condition 3.4.1), while the missing edge between z_{pt} and $Y_{pt}(\overline{\mathbf{a}}_{\mathbf{pt}})$ indicates that the value of the instrumental variable Z_{pt} has no direct influence on student test performance (i.e., condition 3.4.2). The first assumption is testable (and I will verify this condition in section 3.5.3), and I would argue that it is difficult to

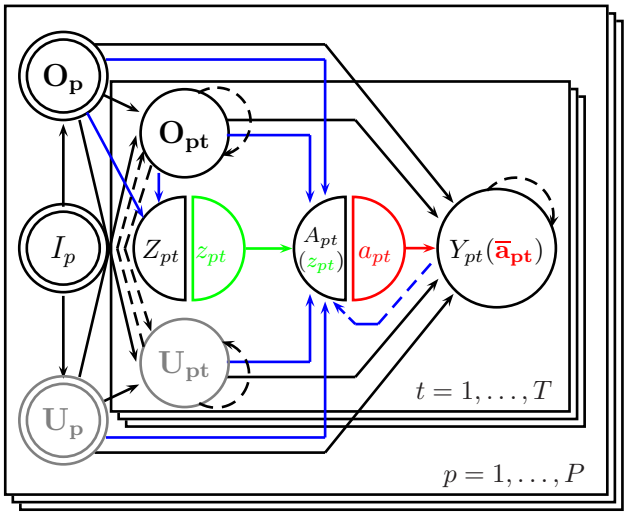


Figure 3.4.2: Substantive assumptions justifying IV model

argue that the second assumption does not hold in this application; i.e., there is no reason to believe that a student's test performance would be impacted *directly* by the district's position relative to the state's special education funding threshold.

The condition that is *not* implied by the SWOOPs in figure 3.4.1, though, is condition 3.4.3. In figure 3.4.2, I illustrate the substantive assumptions that imply condition 3.4.3 ($Z_{pt} \perp\!\!\!\perp Y_{pt}(a_{pt}), A_{pt}(z_{pt}) \mid \mathbf{X_{pt}}$ for $a_{pt}, z_{pt} \in \{0, 1\}$ and $\forall\ p, t$). There are two important aspects of the SWOOPs in figure 3.4.2. First, none of the edges into $A_{pt}(z_{pt})$ are missing, which implies that an IV analysis does not rely on *any untestable assumptions* about how students are selected for special education. Second, the edges between $\mathbf{U_p}$ and Z_{pt} and between $\mathbf{U_{pt}}$ and Z_{pt} are both missing. This implies that an IV analysis *does* rely on the substantive (and not directly testable) assumption that a district's position relative to the state's special education funding threshold is not influenced by any *unobserved* attributes that *also influence student test performance*.

Two aspects of the analysis in section 3.5 are intended to make these substantive assumptions more plausible. Because the percent of special education students in the district clearly influences Z_{pt} and may also impact the performance of students in the district, I include a quadratic of this variable in the vector $\mathbf{X_{O_{pt}}}$. Further, because districts that allow their special education rates to cross the state's funding threshold may be different in unobserved ways from other districts in ways that influences student test performance, I standardize student test scores *within districts*, so that each student's test performance (the outcome) is measured relative to other students in the same district. Having made these two corrections, I would argue that the substantive assumptions illustrated in figure 3.4.1 are plausible in this application. I thus proceed with an IV analysis in section 3.5, but further contrast the assumptions that justify this approach with the assumptions that justify methods in the

existing literature (Hanushek et al., 2002; Morgan et al., 2010) in section 3.6.

3.5 IV Analysis

3.5.1 Data for IV Analysis

To carry out the IV analysis, I make several additional changes to the analytic dataset described in section 3.2.3 (I also perform a second version of the replication study from section 3.2.4 using this dataset in section 3.5.2). Most importantly, I drop all students who were already receiving special education services by the end of the first month of each school year (September). I do this to isolate the influence of the funding threshold from other confounding factors. Specifically, not only can districts not control the movement of students into and out of the district over the summer, but districts may not know their special education enrollment rate (and therefore may not be able to respond to the funding threshold) until the end of the September, when they first need to report this rate to the state for funding purposes. I also drop all observations from the 2009-10 school year, because I need lagged special education data to ensure that students were not receiving special education services prior to the end of September. Thus the IV analysis below is based on data from 2010-11 through 2012-13. Finally, since I hypothesize that large districts will both be more able to and have more incentive to respond to the special education funding threshold, I only include students from the 32 Washington school districts with at least $10,000$ students.

At the end of September of each school year, there are two types of school districts in the state. Districts that have fewer than 12.7% of their students enrolled in special education will continue to receive additional funds from the state for each student in the district that gets placed in special education. But districts that are already beyond the funding threshold (i.e., have more than 12.7% of their students enrolled in special education) will not receive any additional special education funding from the state regardless of the number of students placed in special education over the rest of the year. The instrumental variable Z_{pt}, then, is an indicator for whether each student in the final sample attended a district that was already beyond the state special education funding threshold by the end of September. My hypothesis is that students in districts that are already beyond the funding threshold should be less likely to be placed in special education, all else equal, than comparable students in districts that are not beyond the funding threshold. I investigate this hypothesis within a special education placement model in section 3.5.3.

The dataset for the IV analysis ("IV dataset") contains 454,919 student/year observations in which the student was not enrolled in special education by the end of September of the school year. Of these observations, 1,685 were placed into special education for a specific learning disability over the course of the school year. These students are the treatment group in this study, while the other 453,234 student/year observations serve as the control group. Not surprisingly, the average student in the treatment group scored more than a standard deviation below the mean both at the end of the previous year (-1.17 in reading, -1.11 in math) and at the end of the current year (-1.13 in reading, -1.07 in math). The small

gains in performance for students in the treatment group (.04 standard deviations of student performance in both reading and math) are statistically significant at the .05-level in both subjects, meaning that – without controlling for any confounders – students in the treatment group scored higher (relative to their peers) after receiving special education services for a specific learning disability than before.

The IV analysis relies on detailed data about the percent of students in the district who are enrolled in special education. While I could calculate a rough measure of these percentages from the student-level data, the percentages that really matter are those used by the state for funding purposes. Fortunately, the OSPI posts two annual reports for each district in the state with school funding data specific to special education. The first, Report 1735T, contains the monthly special education enrollment of the district, as reported to the state, along with the average special education enrollment for the months October-June. This average special education enrollment then appears on the second report, Report 1220F, which calculates the "Special Education Excess Cost Allocation" for each district (i.e., the additional funding each district receives from the state to provide special education services). As an example, this funding report for the Spokane School District in the 2011-12 school year demonstrates that the funding threshold cost the district almost $1.3 million in special education funding in this year alone. This provides further justification for the IV approach by reinforcing the point that districts have a considerable incentive to respond to the state's threshold in special education funding.

3.5.2 Replication Study with IV Dataset

Before describing the results from the IV analysis in section 3.5.3, I first perform a second version of the replication study from section 3.2.4 using the IV dataset described in section 3.5.1. Not only will this version of the replication study allow me to compare estimates from all three methods–student fixed effects, covariate adjustment, and instrumental variables–that are estimated from the same dataset, but this version of the replication study is interesting in its own right. Specifically, while the estimates reported in section 3.2.4 are informed by different special education students – i.e., the estimates from the covariate adjustment model are informed by all special education students, while the estimates from the student fixed effects model are informed only by students who transition into and out of special education – the estimates in this version of the replication study are informed by the *exact same students*. This is because of the data restrictions I describe in section 3.5.1 that ensure that every special education student in the sample was *not* receiving special education services in the prior school year. The estimates from this second replication study are in table 3.5.1.

Despite the additional data restrictions, it is perhaps surprising that the differences between the estimates from the student fixed effects model and covariate adjustment model are just as extreme than they were in the first replication study. The primary difference between the estimates from the two replication studies is that the estimated ATE of special education services on student performance from the student fixed effects model estimated from the IV dataset is no longer statistically significant in either reading or math. Given that

	Reading		Math	
	SFE	**COV**	**SFE**	**COV**
Intervention (A_{pt})				
LD services	0.017	-0.254***	0.001	-0.220***
	(0.029)	(0.015)	(0.025)	(0.014)
Time-variant covariates ($\mathbf{O_{pt}}$)				
FRL eligibility	0.001	-0.090***	-0.000	-0.081***
	(0.007)	(0.002)	(0.006)	(0.002)
Gifted services	-0.044***	0.152***	-0.068***	0.201***
	(0.008)	(0.004)	(0.007)	(0.003)
ELL services	0.072***	-0.211***	-0.016	-0.049***
	(0.011)	(0.004)	(0.010)	(0.004)
Migrant services	-0.033	0.013	0.018	0.057***
	(0.025)	(0.010)	(0.022)	(0.009)
Homeless services	-0.024	-0.065***	-0.026	-0.073***
	(0.018)	(0.009)	(0.015)	(0.009)
Time-invariant covariates ($\mathbf{O_p}$)				
Female		0.145***		-0.025***
		(0.002)		(0.002)
American Indian		-0.088***		-0.065***
		(0.011)		(0.011)
Asian / Pacific Islander		0.029***		0.126***
		(0.003)		(0.003)
Black		-0.037***		-0.083***
		(0.004)		(0.004)
Hispanic		-0.022***		-0.031***
		(0.003)		(0.003)
Controls for prior test scores	No	Yes	No	Yes
Student fixed effect	Yes	No	Yes	No

*$p < .05$; **$p < .01$; ***$p < .001$

Table 3.5.1: Estimated coefficients from replication study on IV dataset

these estimates are informed by the exact same students, this replication study provides a particularly striking example of Lord's Paradox; in fact, I used the summary statistics from this replication study in the application in chapter 1.

3.5.3 IV Estimates

Placement Model

As the first step of the IV analysis, I estimate a logistic regression model in which the response variable is the log odds that an individual student is placed into special education for a specific learning disability, and predictor variables are selected through an extensive model selection procedure that uses the Bayesian Information Criterion (BIC, Schwarz et al., 1978) to determine whether the addition of an additional covariate improves the fit of the model (relative to the number of parameters in the model). The final set of predictor variables includes a cubic of prior test scores (standardized both within and across districts), the full vector of student attributes (race, gender, and program participation), a quadratic of district percent special education students at the end of September, an indicator for the special education funding threshold, and grade and year effects.

Estimates from the best-fitting placement model are reported in the first column of Table 3.5.2. All estimates are on the log odds scale, and can be interpreted as the expected change in the log odds of special education placement for a specific learning disability for each unit change in the predictor variable. Some interesting tangential findings from this model are that students in the LEP (Limited English Proficiency) program, female students, and Asian/Pacific Islander students are all less likely to be placed into special education for a specific learning disability (relative to non-LEP, male, and white students, respectively), all else equal. Most importantly, though, the log odds that a student in a district that is beyond the funding threshold gets placed into special education for a specific learning disability is 0.681 less, all else equal, than the log odds for student in a district that is not beyond the funding threshold. That is, even controlling for other predictors of special education placement, the funding threshold does appear to be highly predictive of special education placement in the expected direction. Moreover, the F-statistic on this instrumental variable ($F \approx 30$) is far larger than the recommended minimum ($F = 10$) to avoid problems with weak instruments in IV analyses (e.g., Bound et al., 1995).

Figure 3.5.1 shows predicted probabilities of special education placement (from the placement model reported in column 1 of Table 3.5.2) for four hypothetical students as a function of the percent of special education students in the district. Each hypothetical student is white, male, and is not enrolled in any special programs. The differences between the students are grade level (4th grade or 6th grade) and prior-year test performance (2 standard deviations or 1 standard deviation below the mean in both math and reading). For each student, the predicted probability of placement into special education for a specific learning disability increases as the percent of special education students in the district increases, but drops sharply as the percent of special education students crosses the funding threshold (12.7%). This illustrates the magnitude of the estimated effect of the funding threshold on

special education placement. For example, the predicted probability of special education placement for a white, male 4th grader with low prior performance in both math and reading (2 SDs below the mean) drops by almost ten percentage points as the percent of special education students in the district passes the funding threshold.

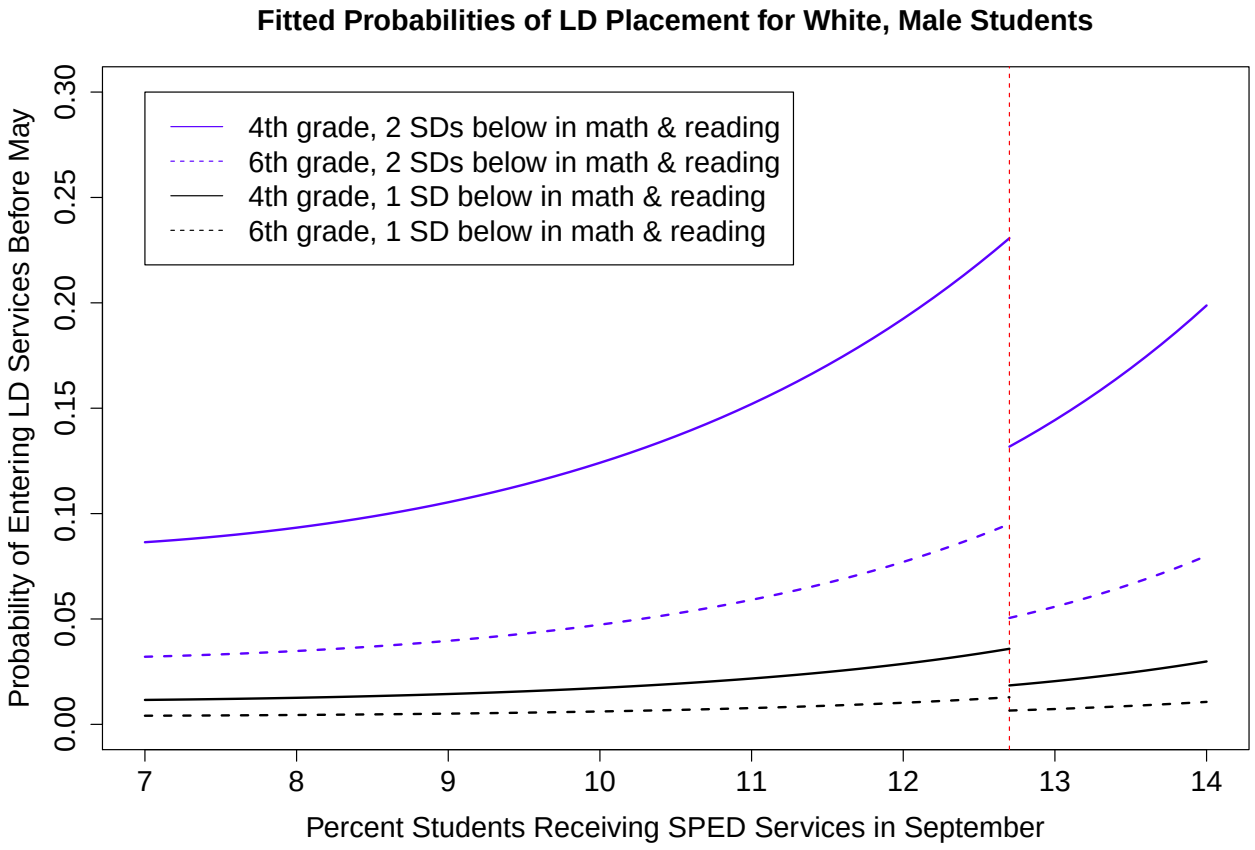


Figure 3.5.1: Fitted probabilities of placement in special education for a specific learning disability

I perform two robustness checks to test the plausibility of these findings. First, as discussed in section 3.2.3, the funding threshold should only influence placement rates into disability categories with subjective placement criteria. When I estimate the same placement model for special education placement into disability types other than specific learning disabilities, I do not find a statistically significant effect for the funding threshold (and the magnitudes of these effects are much smaller than the estimated magnitude for placement into special education for a specific learning disability). This is consistent with the hypothesis that the funding threshold should have a larger impact on student placement into the specific learning disability category than on placement into other categories of special education.

Second, if the funding threshold really influences student placement into special education for a specific learning disability, then it should be the *only* threshold with such an impact (since there are no other thresholds in the special education funding law). That is, I should not observe sharp discontinuities in placement rates at any other value of the percent of special education students in the district. To investigate this, I include other thresholds – 8.7%, 9.7%, 10.7%, 11.7%, and 13.7% – in the placement model, both individually and together, while keeping the 12.7% threshold in the model. None of these other thresholds is a statistically significant predictor of special education placement for a specific learning disability, and the true funding threshold (12.7%) is still a statistically significant predictor in each model (moreover, the coefficient for the true funding threshold is consistently larger than the coefficients for the other thresholds). This is consistent with the hypothesis that the funding threshold is influencing special education placement rates.

Two-Stage Least Squares Model

The primary conclusion from section 3.5.3 – that districts do appear to respond to the funding threshold in special education funding – is important in itself. However, in this section, I use the estimates from the placement model from section 3.5.3 as the first stage of a two-stage least squares (2SLS) model that provides a consistent estimate of the IV estimand discussed in section 3.4.2 (Wooldridge, 2012). I estimate one version of the second-stage model in which the dependent variable is the student's reading score at the end of the year, and another in which the dependent variable is the student's math score at the end of the year.

One technical challenge is that – although the instrumental variable is binary and logistic regression is appropriate for modeling a binary variable – it is not valid to use a non-linear regression (like logistic regression) in the first stage of a 2SLS model (Angrist and Pischke, 2008). So, following the recommendation in Wooldridge (2012), I use the fitted values from the placement model (column 1 of Table 3.5.2) as the instrumental variables in the 2SLS model. A second technical challenge is that the overall rates of special education placement are so low in the IV dataset that the second-stage IV estimates generated from the IV dataset are very imprecise (intuitively, because the estimate of the denominator of β^{IV} is extremely small). To generate more precise estimates in the second-stage model, I match each student who is placed into special education with nine students with identical prior test scores who were not placed into special education, and thus create a "matched" dataset in which exactly 10% of students are placed into special education.

Columns 2 and 3 of Table 3.5.2 show the estimates from the second stage of this model (for reading and math, respectively) estimated from this matched dataset. The coefficient on the variable of interest, "LD services" (i.e., whether or not the student received special education services for a specific learning disability), is not statistically significant in either math or reading. That is, the estimates from the IV model (even from the matched dataset) do not provide sufficient evidence to conclude that the average impact of special education services for a specific learning disability on student performance is different from zero in either subject.

	1st stage model	2nd stage models	
	Placement	**Reading**	**Math**
Intervention (A_{pt})			
LD services		-0.055	0.016
		(0.143)	(0.128)
Instrumental variable (Z_{pt})			
Beyond threshold	-0.681***		
	(0.124)		
Time-variant covariates ($\mathbf{O_{pt}}$)			
FRL eligibility	-0.089	-0.078***	-0.071***
	(0.063)	(0.015)	(0.013)
Gifted services	-0.319	0.119*	0.061
	(0.342)	(0.060)	(0.057)
ELL services	-0.163*	-0.190***	-0.078***
	(0.069)	(0.016)	(0.015)
Migrant services	-0.075	0.046	0.025
	(0.199)	(0.036)	(0.034)
Homeless services	-0.060	-0.094*	-0.053
	(0.154)	(0.040)	(0.033)
Time-invariant covariates ($\mathbf{O_p}$)			
Female	-0.108*	0.107***	-0.023*
	(0.051)	(0.012)	(0.011)
American Indian	0.053	-0.013	-0.071
	(0.221)	(0.056)	(0.048)
Asian / Pacific Islander	-0.631***	0.036	0.064**
	(0.115)	(0.024)	(0.023)
Black	-0.124	-0.071**	-0.083***
	(0.088)	(0.022)	(0.019)
Hispanic	-0.044	-0.003	-0.022
	(0.075)	(0.018)	(0.016)

*$p < .05$; **$p < .01$; ***$p < .001$.

Table 3.5.2: Estimated coefficients from IV analysis

Although the precision of the IV estimates is insufficient to distinguish a significant treatment effect, it is interesting to compare the IV estimates from table 3.5.2 with the estimates from the replication study that uses the same dataset (table 3.5.1). Specifically, the modest estimates from the IV model are much more consistent with the estimates from the student fixed effects model (Hanushek et al., 2002) than with the large, negative estimates from the covariate adjustment model (Morgan et al., 2010). This is broadly consistent with two observations from section 3.3.3: (1) the assumptions that justify a student fixed effects model are perhaps more plausible than the assumptions that justify a covariate adjustment model in this application; and (2) the "omitted variable bias" in the estimates from the covariate adjustment model is likely to be negative.

3.6 Conclusions

The overarching conclusions from this chapter largely parallel the conclusions from the investigation of Lord's Paradox in chapter 1; namely, estimating causal effects from observational data *always* requires untestable assumptions, so the key to any observational study is finding an approach (if it exists!) that is justified by assumptions that are plausible in the given application. In this chapter, I have discussed three approaches to estimating the ATE of special education services on student performance – a student fixed effects model, a covariate adjustment model, and an instrumental variables (IV) model – that are justified by very different assumptions:

- A student fixed effects model is justifiable if students are not selected for special education on the basis of prior test performance or any unobserved variables that change over time and influence student test performance.
- A covariate adjustment model is justifiable if students are not selected for special education on the basis of any unobserved variables that influence student test performance.
- An IV model is justifiable if a district's position relative to the state's special education funding threshold is not a function of any unobserved variables that also influence student test performance.

In this chapter, I have argued that the substantive assumptions justifying an IV model are more plausible than the assumptions justifying a student fixed effects model, which in turn are more plausible than the assumptions justifying a covariate adjustment model. But reasonable researchers could make arguments supporting *any* of these assumptions, which leads to the second broad conclusion: the graphical framework developed in chapter 2 and used throughout this chapter (SWOOPs) provide a simple and intuitive way of communicating these assumptions and verifying that they imply the counterfactual conditions justifying each method. Given the prevalence of observational studies in education policy research and the broader social sciences, I believe that more widespread use of graphs (and SWOOPs in particular) would help researchers communicate and clarify the assumptions that justify causal conclusions from this vast line of research.

Bibliography

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434):444–455.

Angrist, J. D. and Pischke, J.-S. (2008). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Bangsø, O. and Wuillemin, P.-H. (2000). Object oriented Bayesian networks: A framework for top-down specification of large Bayesian networks with repetitive structures.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association, 90(430):443–450.

Buntine, W. L. (1994). Operations for learning with graphical models. arXiv preprint cs/9412102.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2013). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. Technical report, National Bureau of Economic Research.

Cullen, J. B. (2003). The impact of fiscal incentives on student disability rates. Journal of Public Economics, 87(7):1557–1589.

Dawid, A. P., Mortera, J., and Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic dna profiling problems. Forensic Science International, 169(2):195–205.

Dhuey, E. and Lipscomb, S. (2011). Funding special education by capitation: Evidence from state finance reforms. Education Finance and Policy, 6(2):168–201.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(1):245–264.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. Journal of Econometrics, 137(2):334–353.

Ewing, K. A. (2009). Estimating the Effectiveness of Special Education Using Large-Scale Assessment Data. PhD thesis, North Carolina State University.

Geiger, D. (1990). Graphoids: A qualitative framework for probabilistic inference. PhD thesis, University of California at Berkeley.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in Bayesian networks. Networks, 20(5):507–534.

Gelman, A. and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., and Max, J. (2013). Transfer incentives for high-performing teachers: Results from a multisite randomized experiment. Technical report, Mathematica Policy Research.

Goldhaber, D. and Theobald, R. (2013a). Do different value-added models tell us the same things? Carnegie Knowledge Network Briefs. Stanford, CA.

Goldhaber, D. and Theobald, R. (2013b). Managing the teacher workforce in austere times: The implications of teacher layoffs. Education Finance and Policy, 8(4):494–527.

Goldstein, L. (2003). Spec. ed. growth spurs cap plan in pending idea. Education Week, 22(1):16–17.

Greene, J. P. and Forster, G. (2002). Effects of funding incentives on special education enrollment. Technical report, Manhattan Institute for Policy Research.

Hanushek, E. A., Kain, J. F., and Rivkin, S. G. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? Review of Economics and Statistics, 84(4):584–599.

Heckerman, D., Meek, C., and Koller, D. (2007). Probabilistic entity-relationship models, prms, and plate models. Introduction to Statistical Relational Learning, pages 201–238.

Holland, P. and Rubin, D. (1983). On Lord's paradox. In Wainer, H. and Messick, S., editors, Principles of Modern Psychological Measurement, pages 3–25. Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960.

Hong, G. and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. Journal of the American Statistical Association, 101(475).

IDEA (1975). Individuals with disabilities education act.

IDEA (2004). Individuals with disabilities education act.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. Econometrica, 62(2):467–475.

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. Review of Economics and Statistics, 86(1):226–244.

Jamieson, J. (1999). Dealing with baseline differences: Two principles and two dilemmas. International Journal of Psychophysiology, 31(2):155–161.

Kane, D. and Johnson, P. (1993). Vermont's Act 230: A new response to meeting the demands of diversity. Vermont Department of Education, Montpelier, VT.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research paper. MET Project. Bill & Melinda Gates Foundation.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence, pages 302–313. Morgan Kaufmann Publishers Inc.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. Networks, 20(5):491–505.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. Psychological Bulletin, 68(5):304–305.

Lord, F. M. (1975). Lord's paradox. In Anderson, S. B., Ball, S., and Murphy, R. T., editors, Encyclopedia of Educational Evaluation, pages 232–236. Jossey-Bass Publishers, San Francisco.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modeling framework: concepts, structure, and extensibility. Statistics and Computing, 10(4):325–337.

Mahitivanichcha, K. and Parrish, T. (2005). Do non-census funding systems encourage special education identification?: Reconsidering Greene and Forster. Journal of Special Education Leadership, 18(1):38–46.

Maxwell, S. E. and Delaney, H. D. (2004). Designing Experiments and Analyzing Data: A Model Comparison Perspective. Psychology Press.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29(1):67–101.

Morgan, P. L., Frisco, M. L., Farkas, G., and Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. The Journal of Special Education, 43(4):236–254.

Murphy, K. P. (2002). Dynamic Bayesian Networks: Representation, Inference and Learning. PhD thesis, University of California, Berkeley.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych, 10:1–51.

Nodelman, U., Shelton, C. R., and Koller, D. (2002). Continuous time Bayesian networks. In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, pages 378–387. Morgan Kaufmann Publishers Inc.

Nye, B., Hedges, L. V., and Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. American Educational Research Journal, 37(1):123–151.

Ogburn, E. L. and VanderWeele, T. J. (2014). Causal diagrams for interference. Statistical Science, 29(4):559–578.

Parker, J. M. (2011). The Use of Large-Scale Assessment Data to Investigate Special Education Effectiveness. PhD thesis, North Carolina State University.

Pearl, J. (1995). Causal diagrams for empirical research. Biometrika, 82(4):669–688.

Pearl, J. (2000). Causality: Models, Reasoning and Inference, volume 29. Cambridge Univ Press.

Pearl, J. (2014). Lords paradox revisited–(oh Lord! kumbaya!). Technical report, University of California - Los Angeles.

Rattigan, M. J. and Jensen, D. (2010). Leveraging d-separation for relational data sets. In ICDM, pages 989–994.

Rattigan, M. J., Maier, M. E., and Jensen, D. (2011). Relational blocking for causal discovery. In AAAI.

Raudenbush, S. W. and Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods, volume 1. Sage.

Reardon, S. F. and Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. Education Finance and Policy, 4(4):492–519.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical report, University of Washington CSSS Working Paper 128.

Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., and Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. Frontiers in Artificial Intelligence and Applications, 162:13.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. American Economic Review, pages 247–252.

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

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. Education Finance and Policy, 4(4):537–571.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. The Annals of Statistics, pages 34–58.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. Journal of the American Statistical Association, 81(396):961–962.

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 29:103–116.

Schwarz, G. et al. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). Causation, Prediction, and Search, volume 81. MIT press.

Wainer, H. (1991). Adjusting for differential base rates: Lord's paradox again. Psychological Bulletin, 109(1):147.

Wooldridge, J. M. (2012). Introductory Econometrics: A Modern Approach. Cengage Learning.

Wright, D. B. (2006). Comparing groups in a before–after design: When t-test and ANCOVA produce different results. British Journal of Educational Psychology, 76(3):663–675.

Appendix A

Proof of Markov property in figure 1.3.2b

First notice that the joint distribution of the variables in figure 1.3.2b can be directly related to the joint distribution of the variables in figure 1.3.2a:

$$\begin{aligned} & \Pr(\Delta Y_2(a_2) = \delta, M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2^*, Y_1 = y_1) \\ = \; & \Pr(Y_2(a_2) - Y_1 = \delta, M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2^*, Y_1 = y_1) \\ = \; & \Pr(Y_2(a_2) = \delta + y_1, M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2^*, Y_1 = y_1). \end{aligned} \tag{A.0.1}$$

Under the assumption that the distribution P is Markov with respect to the SWIT in figure 1.3.2a, equation A.0.1 can be written as:

$$\begin{aligned} = \; & \Pr(Y_2(a_2) = \delta + y_1 | M = m, U_2 = u_2, A_2 = a_2) \\ & \times \Pr(A_2 = a_2^* | M = m) \times \Pr(Y_1 = y_1 | M = m, U_1 = u_1) \\ & \times \Pr(U_2 = u_2 | U_1 = u_1) \times \Pr(M = m) \times Pr(U_1 = u_1) \text{ by definition 1.2.4.} \end{aligned} \tag{A.0.2}$$

We can now re-write the first term in equation A.0.2:

$$\begin{aligned} = \; & \Pr(Y_2(a_2) = \delta + y_1 | M = m, U_2 = u_2, A_2 = a_2) \\ = \; & \Pr(Y_2(a_2) = \delta + y_1 | M = m, U_1 = u_1, U_2 = u_2, A_2 = a_2) \text{ by equation 1.3.3} \\ = \; & \int_{y_1} \Pr(Y_2(a_2) = \delta + y_1, Y_1 = y_1 | M = m, U_1 = u_1, U_2 = u_2, A_2 = a_2) dy_1 \\ = \; & \int_{y_1} \Pr(Y_2(a_2) - Y_1 = \delta, Y_1 = y_1 | M = m, U_1 = u_1, U_2 = u_2, A_2 = a_2) dy_1 \\ = \; & \int_{y_1} \Pr(\Delta Y_2(a_2) = \delta, Y_1 = y_1 | M = m, U_1 = u_1, U_2 = u_2, A_2 = a_2) dy_1 \\ = \; & \Pr(\Delta Y_2(a_2) = \delta | M = m, U_1 = u_1, U_2 = u_2, A_2 = a_2) \\ = \; & \Pr(\Delta Y_2(a_2) = \delta | U_1 = u_1, U_2 = u_2, A_2 = a_2) \text{ by equation 1.3.6.} \end{aligned}$$

Going back to equations A.0.1 and A.0.2, this means that the joint distribution of the variables $\mathbf{V}$ in figure 1.3.2b can be written as:

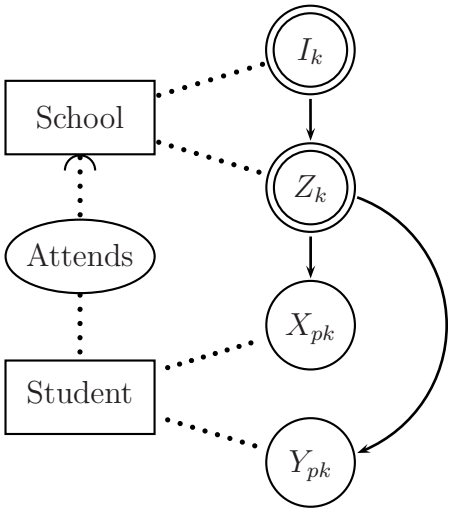
$$\begin{aligned} P(\mathbf{V}) &= \Pr(\Delta Y_2(a_2) = \delta, M = m, \mathbf{U} = \mathbf{u}, A_2 = a_2, Y_1 = y_1) \\ &= \Pr(\Delta Y_2(a_2) = \delta | U_1 = u_1, U_2 = u_2, A_2 = a_2) \\ &\quad \times \Pr(A_2 = a_2^* | M = m) \times \Pr(Y_1 = y_1 | M = m, U_1 = u_1) \\ &\quad \times \Pr(U_2 = u_2 | U_1 = u_1) \times \Pr(M = m) \times Pr(U_1 = u_1) \\ &= \prod_{V \in \mathbf{V}} P(V | pa_{\mathcal{G}}(V)) \text{ for the SWIT } \mathcal{G} \text{ in figure 1.3.2b.} \end{aligned} \tag{A.0.3}$$

So by definition 1.2.4, we can conclude that if the distribution P is Markov with respect to the SWIT in figure 1.3.2a, it must also be Markov with respect to the SWIT in figure 1.3.2b.

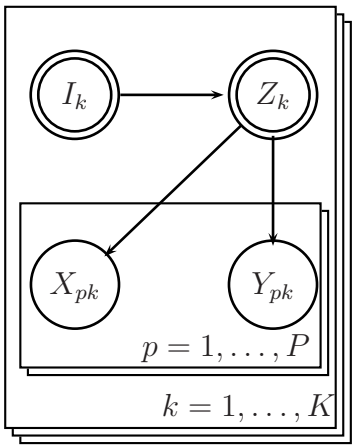
Appendix B

Discussion of ID variables (Rattigan et al., 2011)

Each application of ID variables in chapter 2 is analogous to the example of a *school-level confounder* from Rattigan et al. (2011) (figure 2b), where I_k is an ID variable for school k, and Z_k is a school-level attribute that influences individual-level attributes X_{pk} and Y_{pk}. We show the corresponding DAPER diagram and stacked exchangeable plates in figure B.0.1.



(a) DAPER diagram



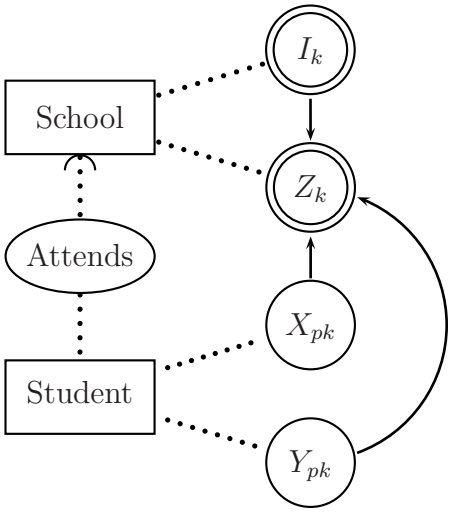
(b) Stacked exchangeable plates

Figure B.0.1: School-level confounder (Rattigan et al., 2011, figure 2b)

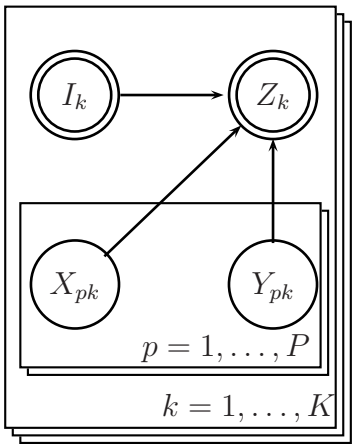
We view this application of ID variables as a natural and intuitive extension of DAGs and DAPER graphs, as it is clear that the school-level attribute Z_k is a deterministic attribute in that its only parent is the ID variable I_k (and both variables can be regarded as *fixed*, as emphasized by the double-circle around each node). Moreover, the stacked exchangeable plates in figure B.0.1b allow us to verify the well-known result (discussed extensively in Rattigan et al. (2011)) that controlling for a school-level confounder or stratifying by school level (i.e., controlling for the ID variable I_k) are both sufficient to control for school-level confounding.

Specifically, while $X_{pk} \not\perp\!\!\!\perp Y_{pk}$ in figure B.0.1b, the rules of D-separation (definition 2.2.9) allow us to conclude that $X_{pk} \perp\!\!\!\perp Y_{pk}|Z_k$ and $X_{pk} \perp\!\!\!\perp Y_{pk}|I_k$ in figure B.0.1b.

However, Rattigan et al. (2011) also apply ID variables to the case of a *school-level collider*, and we do *not* believe that this application of ID variables is justifiable. In their example, Rattigan et al. (2011) define $X'_k = \sum_{p=1}^{P} X_{pk}$ and $Y'_k = \sum_{p=1}^{P} Y_{pk}$, and let $Z_k = \beta X'_k + \beta Y'_k + \epsilon_k$ for some $\beta \neq 0$. Figure B.0.2 illustrates the graphical representation of this example from Rattigan et al. (2011) as a DAPER diagram and as stacked exchangeable plates.



(a) DAPER diagram



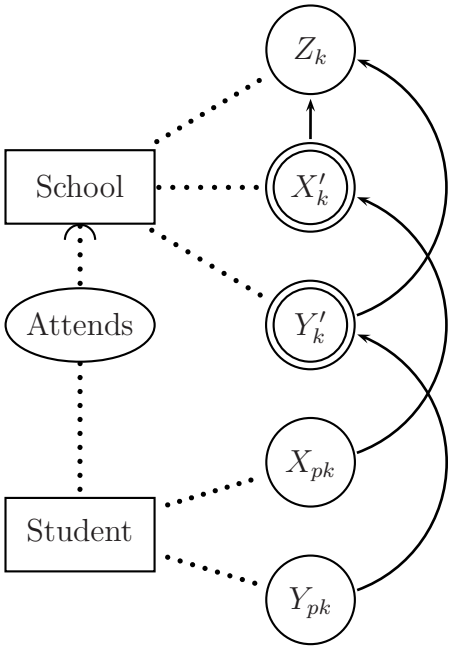
(b) Stacked exchangeable plates

Figure B.0.2: School-level collider (Rattigan et al., 2011, figure 2d)

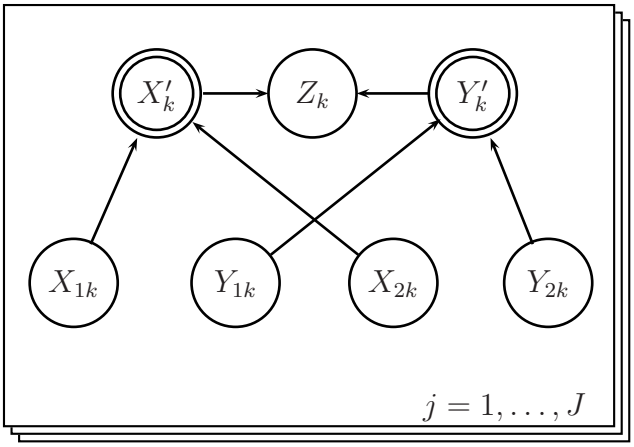
As noted by Rattigan et al. (2011), the utility of the graph in figure B.0.2b is that we can apply the rules of D-separation to conclude that, in this example, $X_{pk} \not\perp\!\!\!\perp Y_{pk}|Z_k$ but $X_{pk} \perp\!\!\!\perp Y_{pk}|I_k$. While this leads to the correct conclusion that controlling for a school-level confounder introduces bias while stratifying by school-level does not in this example, we do not believe that the graph in figure B.0.2b is internally consistent. Specifically, Z_k can no longer be regarded as fixed if it is a child of X_{pk} and Y_{pk}, so it is unclear how it can be a deterministic attribute. Thus we intentionally avoid this application of ID variables in chapters 2 and 3.

That said, we can draw the same conclusions from Rattigan et al. (2011) without using ID variables. Figure B.0.3 illustrates an alternative approach for the case where $P = 2$. Note that we have included the school-level sums $X'_k = \sum_{p=1}^{P} X_{pk}$ and $Y'_k = \sum_{p=1}^{P} Y_{pk}$ as deterministic attributes on the graph, and included Z_k as a child of these school-level sums. Applying the rules of D-separation to figure B.0.3b leads to the conclusions that $X_{pk} \not\perp\!\!\!\perp Y_{pk}|Z_k$ but $X_{pk} \perp\!\!\!\perp Y_{pk}|X'_k, Y'_k$. The first conclusion is identical to Rattigan et al. (2011), while the second conclusion justifies stratifying by school level in this example because including a school-effect in a model is equivalent to controlling for school-level means of the individual-

level variables (Angrist and Pischke, 2008). Thus we argue that the *conclusion* in Rattigan et al. (2011) is correct, but that a graph without ID variables offers a better way to illustrate and verify this conclusion in a graphical framework.



(a) DAPER diagram



(b) Stacked exchangeable plates

Figure B.0.3: An alternative representation of Rattigan et al., 2011, figure 2D

Appendix C

Monotonicity and Local Average Treatment Effects (Imbens and Angrist, 1994)

In section 3.4.2 of chapter 3, I make the following constant treatment effect assumption:

Condition C.0.1 $Y_{pt}(1) - Y_{pt}(0) = E(Y_{pt}(1) - Y_{pt}(0)) \; \forall \; p, t.$

But it is very difficult to argue that condition C.0.1 is realistic in this application; i.e., not all students will respond to special education services in exactly the same way. Thus the derivation in chapter 3 that the IV estimator $\beta^{IV} = \mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$ relies on an assumption that probably does not hold. However, Imbens and Angrist (1994) substitute condition C.0.1 for a weaker condition that I argue is far more plausible in this application:

Condition C.0.2 *(Monotonicity)* $A_{pt}(1) \leq A_{pt}(0) \; \forall \; p, t.$

This condition states that every student who *would* have been placed in special education in a district beyond the funding threshold also would have been placed in special education if the district was not beyond the funding threshold, and every student who *would not* have been placed in special education in a district not beyond the funding threshold also would not have been placed in special education if the district was beyond the funding threshold. I argue that this is a much more plausible assumption, as it is difficult to argue why the funding threshold would ever create a "reverse incentive" to either identify or not identify a specific student for special education services.

Under condition C.0.2 and conditions 3.4.1-3.4.3 from chapter 3, Imbens and Angrist (1994) show that β^{IV} is equal to the Local Average Treatment Effect (LATE) $E(Y_{pt}(1) - Y_{pt}(0) | A_{pt}(1) \neq A_{pt}(0))$ (this is proven on the following page). This is a much more specific estimand than the ATE $\mathrm{E}(Y_{pt}(1) - Y_{pt}(0))$, as it is defined only for students whose treatment assignment is influenced by the value of the instrument. But it can be identified under assumptions that are much more plausible than the assumptions outlined in chapter 3.

$$
\begin{aligned}
& \mathrm{E}\{\mathrm{E}(Y_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(Y_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \;& \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)+(1-A_{pt}(1))Y_{pt}(0)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& -\mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)+(1-A_{pt}(0))Y_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& \text{by condition 3.4.2} \\
= \;& \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& +\mathrm{E}\{(1-A_{pt}(1))Y_{pt}(0)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& -\mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& +\mathrm{E}\{(1-A_{pt}(0))Y_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
= \;& \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& +\mathrm{E}\{(1-A_{pt}(1))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& -\mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\} \\
& +\mathrm{E}\{(1-A_{pt}(0))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \text{ by condition 3.4.3} \\
= \;& \mathrm{E}\{\mathrm{E}\{A_{pt}(1)Y_{pt}(1)+(1-A_{pt}(1))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
& -\mathrm{E}\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)+(1-A_{pt}(0))Y_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}}\}\} \\
= \;& \mathrm{E}\{A_{pt}(1)Y_{pt}(1)+(1-A_{pt}(1))Y_{pt}(0)\} \\
& -\{\mathrm{E}\{A_{pt}(0)Y_{pt}(1)+(1-A_{pt}(0))Y_{pt}(0)\} \\
= \;& \mathrm{E}\{(A_{pt}(1)Y_{pt}(1)-A_{pt}(1)Y_{pt}(0)-A_{pt}(0)Y_{pt}(1)+A_{pt}(0)Y_{pt}(0)\} \\
= \;& \mathrm{E}\{(A_{pt}(1)-A_{pt}(0))(Y_{pt}(1)-Y_{pt}(0))\} \\
= \;& \sum_{a_{pt}\in(-1,0,1)} a_{pt}\Pr(A_{pt}(1)-A_{pt}(0)=a_{pt}) \\
& \times \mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)-A_{pt}(0)=a_{pt}) \\
= \;& Pr(A_{pt}(1)-A_{pt}(0)=1)\mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)-A_{pt}(0)=1) \\
& -Pr(A_{pt}(1)-A_{pt}(0)=-1)\mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)-A_{pt}(0)=-1) \\
= \;& -Pr(A_{pt}(1)\neq A_{pt}(0))\mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)\neq A_{pt}(0)) \\
& \text{by condition C.0.2 (since } Pr(A_{pt}(1)-A_{pt}(0)=1)=0) \\
= \;& \mathrm{E}(A_{pt}(1)-A_{pt}(0))\mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)\neq A_{pt}(0))
\end{aligned}
$$

$$
\begin{aligned}
& \mathrm{E}\{\mathrm{E}(A_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(A_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \;& \mathrm{E}\{\mathrm{E}(A_{pt}(1)|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(A_{pt}(0)|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \;& \mathrm{E}\{\mathrm{E}(A_{pt}(1)|\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(A_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \text{ by condition 3.4.3} \\
= \;& \mathrm{E}\{\mathrm{E}(A_{pt}(1)-A_{pt}(0)|\mathbf{X_{pt}}=\mathbf{x_{pt}})\} \\
= \;& \mathrm{E}(A_{pt}(1)-A_{pt}(0))\neq 0 \text{ by condition 3.4.1}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \beta^{IV} \;&= \frac{\mathrm{E}\{\mathrm{E}(Y_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(Y_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\}}{\mathrm{E}\{\mathrm{E}(A_{pt}|Z_{pt}=1,\mathbf{X_{pt}}=\mathbf{x_{pt}})-\mathrm{E}(A_{pt}|Z_{pt}=0,\mathbf{X_{pt}}=\mathbf{x_{pt}})\}} \\
&= \frac{\mathrm{E}(A_{pt}(1)-A_{pt}(0))\mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)\neq A_{pt}(0))}{\mathrm{E}(A_{pt}(1)-A_{pt}(0))} \\
&= \mathrm{E}(Y_{pt}(1)-Y_{pt}(0)|A_{pt}(1)\neq A_{pt}(0))
\end{aligned}
$$