Assumptions of Value-Added Models for Estimating School Effects

Sean F. Reardon (corresponding author)

Stanford University

526 Galvez Mall, #520

Stanford, CA 94305

(650) 736-8517

sean.reardon@stanford.edu

Stephen W. Raudenbush

University of Chicago

1126 East 59th St., Room SS 416

Chicago, IL 60637

(773) 834-1904

sraudenb@uchicago.edu

The work reported here was supported by funds from the William T. Grant Foundation (Reardon) and the Spencer Foundation (Raudenbush) for the project "Improving Research on Instruction: Models, Designs, and Analytic Methods." We thank Derek Neal for his thoughtful discussion of earlier stages of this work. All errors remain our own.

Abstract

The ability of school (or teacher) value-added models to provide unbiased estimates of school (or teacher) effects rests on a set of assumptions. In this paper, we identify six assumptions that are required in order that the estimands of such models are well-defined and that the models are able to recover the desired parameters from observable data. These assumptions are 1) manipulability; 2) no interference between units; 3) interval scale metric; 4) homogeneity of effects; 5) strongly ignorable assignment; and 6) functional form. We discuss the plausibility of these assumptions and the consequences of their violation. In particular, because the consequences of violations of the last three assumptions have not been assessed in prior literature, we conduct a set of simulation analyses to investigate the extent to which plausible violations of them alter inferences from value-added models. We find that modest violations of these assumptions degrade the quality of value-added estimates, but that models that explicitly account for heterogeneity of school effects are less affected by violations of the other assumptions.

1. Introduction

The "No Child Left Behind" Act of 2001 requires states to hold schools accountable for student achievement in several subject areas. Specifically, a school's evaluation depends upon the proportion of its children who score at or above a threshold on a test in each subject area. Such children are proclaimed "proficient" while children scoring below the threshold are deemed "not proficient." A common criticism is that this approach is unfair to schools whose children enter the school with comparatively low levels of cognitive skill. Such schools are less likely to display high proficiency rates later on than are schools serving children whose initial skills are high -- even if all schools are equally effective at promoting learning.

A popular alternative is to hold schools accountable for their "value added." Intuitively, the value added approach holds a school accountable for how much children learn while under the care of that school. A school serving disadvantaged students is not penalized as long as those students learn at a good rate while attending that school. The notion is that if School A produces more learning than does School B during a given year, School A should be ranked higher than School B in a fair accountability system.

Despite the obvious appeal of the value added idea, the precise specification and estimation of value added models is not straightforward. The assertion that School A produces more learning than does School B is, in essence, a causal inference that can be subjected to scrutiny using methods of causal analysis applied more generally in the social sciences. Causal inferences inevitably entail assumptions. To scrutinize a causal inference is to scrutinize those assumptions, requiring that those assumptions be made explicit. The question is not simply whether key assumptions hold, but whether plausible departures from those assumptions would lead to substantially distorted inferences.

Our aim in this paper is therefore to explicate the assumptions that must be made to support value added inferences, to consider the plausibility of those assumptions in practice, and therefore to

lay a strong foundation for evaluating how consequential the violations of those assumptions are likely to be in practical application.

Any explicit strategy for drawing causal inferences will entail two models. The first is a theoretical model that defines the causal effects of interest, and this model is based on certain assumptions that must be specified explicitly if the theory is to be made accessible to criticism. The second is a model for the observable data. Under specific additional assumptions, the parameters of the observed-data model are equivalent to the parameters of the theoretical model for the causal effects. These additional assumptions are said to "identify" the causal effects defined in the theoretical model. Having defined the observed data model and the assumptions that identify it, one must adopt a method for estimating the parameters of the observed data model from a finite sample of data. The estimates should be accurate enough to be of use based on the available data if the enterprise is to succeed. The estimation procedure itself may require additional assumptions, and these should also be made explicit.

1.1 Defining the Causal Effects

Following Neyman (1923/1990), Fisher (1935), Rubin (1978), Heckman (1979), and Holland (1986), Part I of our paper conceptualizes causal effects as person-specific comparisons between potential outcomes associated with alternative treatments. The alternative treatments of interest here are attendance for a year at one of J schools, yielding a potential outcome y_i^j, the achievement test score of student i in school j, there being J potential outcomes for each student. To compare the effectiveness of different schools is then to compare the distributions of the potential outcomes in those schools. Although seemingly quite general, this scheme requires that each student possess one and only one potential outcome in each school. This seemingly simple requirement entails two distinct assumptions:

(i) that it is theoretically meaningful to define the potential outcome of each student if assigned to each of the J schools, ensuring that each student has at least one potential outcome per school;

(ii) that each student possesses no more than one potential outcome in each school, regardless of the school assignment of the other children in the system.

Assumption (i) is known as the assumption of "manipulability" in the causal inference literature (Rosenbaum & Rubin, 1983); (ii) is the "no interference between units" assumption of Cox (1958), closely linked to the "Stable Unit Treatment Value Assumption" (or SUTVA; Rubin, 1986). In the school effects literature, (ii) implies that there are no "peer effects"—that a student's outcome if assigned to a given school does not depend on which other students are assigned to that same school.

Assumption (i) may seem to be implausible, given the reality of school segregation on the basis of various demographic characteristics of students, including family socio-economic background, ethnicity, linguistic background, and prior achievement. These social forces may effectively ensure that, in practice, some students have no access to certain schools, even if their families are capable of moving a considerable distance, so that for those students, no potential outcome exists in those schools. Consider, for example, the potential outcome of a white student from a wealthy background living in an all-white suburb if that student were assigned to an all-minority inner city school. One might argue that such a school assignment is extremely unlikely and therefore of no theoretical or practical interest in light of current policy options. The near-zero likelihood of *observing* this potential outcome (because of the near-zero likelihood of such a school assignment), however, does not mean that we cannot imagine its existence. Thus, we can define causal estimands that depend on its existence, though they may of little or no policy relevance and inestimable in practice, a point we return to later. For now, we simply note that (i) is logically necessary in order to define causal estimands that depend on comparing the

distributions of all students' potential outcomes in two or more schools.

Assumption (ii) may also be implausible, given evidence of peer effects in education and the likely significance of a school's student composition for the organization and delivery of instruction (Barr & Dreeben, 1988). For example, if one's classmates come to school with advanced knowledge in a subject area, the instruction in that subject area may be quite different from the instruction observable had a different set of peers been assigned to that school, particularly peers with meager prior knowledge. Likewise, if school or classroom size affects student outcomes, then the school assignments of other students will affect a given student's outcome.

Further assumptions are typically added to the theoretical models for causal effects, however, to render the parameters of the model estimable from observable data. Thus, most analysts hope to compare schools by comparing their means, overall or for sub-groups, implying that the quantity of interest is the mean difference in potential outcomes associated with any pair of schools, for the population as a whole or for some well-defined sub-population. Reliance on the mean implies

(iii) that the units of the test score distribution are on an interval scale of social interest.

While it is possible to estimate school effects for sub-groups of students, statistical models for value-added causal effects typically do not allow these school effects to vary across students in relation to student characteristics. In part, this reflects the desire of policy makers to use value-added statistics for the purpose of providing an overall ranking of school performance. The absence of interaction of school effects with student characteristics implicitly assumes

(iv) that the causal effects of schools are invariant as a function of student background.

We shall refer to assumption (iii) as the interval-scale metric assumption, or, for short, the "metric assumption;" (iv) is the assumption of homogeneity of causal effects, or, for short, the "homogeneity" assumption. A weaker version of (iv) is a "monotonicity" assumption: if the mean difference between the value added of schools j and k is positive for children of a given background, it is also positive for children of every other background. Other assumptions may also be added to the theoretical model to facilitate estimation, including assumptions of normality, homogeneity of variance, and independence of errors across schools. We shall not emphasize these, however, though we do make occasional reference to them.

Analysts may not wish to make assumptions (iii) and (iv). Assumption (iii) may be relaxed by adopting a non-parametric approach, one that does not compare means bur rather compares distributions of potential outcomes across schools in some more general way, such as comparing the quantiles of each school's distribution of potential outcomes. While appealing, such an approach may require an inordinately large sample size for accurate estimation. Assumption (iv) may be relaxed through the inclusion of interaction terms or random coefficients into the statistical model to be estimated. The next section considers the problem of estimating the parameters of the theoretical model from the observed data.

1.2 Identifying the Causal Effects: The Model for the Observed Data

The "fundamental problem of causal inference" (Holland, 1986) is that, while each student possesses J potential outcomes, one for each school, only one of those potential outcomes will actually be observed. It is therefore impossible to calculate the causal effect of attending one school relative to another for any student. Yet we could *estimate* the distribution of potential outcomes in each school if students were randomly assigned each year to the schools of interest. In that case, the observed outcomes in each school would be a simple random sample of the distribution of potential outcomes

associated with that school. The mean of that sample would estimate the population mean of potential outcomes without bias; and quantiles of the sample would be unbiased estimates of the quantiles of the distribution of potential outcomes for that school. An assumption that equates parameters estimable from the sample data to the parameters of the theoretical model is known as an "identifying assumption." The key identifying assumption supported by random assignment of students is the assumption that a student's potential outcomes are independent of school assignment, conditional on observed characteristics. This is the assumption of "ignorable treatment assignment" (Rosenbaum & Rubin, 1983).

Of course, students are not assigned randomly to schools, and it becomes essential to impose a stronger assumption if we are to make progress. Advocates of value added models therefore assume, within subsets of students who have the values of pre-assignment covariates collected in a matrix **X**, that school assignment is effectively, if not formally, random. In effect, the value-added proponents are assuming:

*(v) that potential outcomes are independent of school assignment, given **X**.*

This assumption has been called the assumption of "strongly ignorable treatment assignment" (Rosenbaum & Rubin, 1983). It requires that any unobserved student characteristics that predict the potential outcomes are independent of school assignment once the observed pre-assignment characteristics of that student are taken into account. In essence, conditioning on the observables in **X** is sufficient to eliminate confounding. This is the same assumption applied in many non-experimental evaluations of new interventions in medicine, social services, education, and economics.

Even given ignorable assignment, the parameters of a statistical model typically rely on the appropriateness of the functional form of the model and/or the availability of data to fit the model. If

school effects depend on student characteristics, for example, but not all schools contain students with any given set of characteristics, then estimation of the mean potential outcome for students in some schools will rely heavily on the functional form of the model—a statistical model extrapolates from regions with data into regions without data by relying on the estimated parameters of the specified functional form. Because students are unevenly distributed among schools, value-added models depend implicitly on the assumption

(vi) that the functional form of the model correctly specifies the potential outcomes even for types of students who are not present in a given school.

This assumption may be termed either the "functional form" or "common support" assumption. It says that either there is adequate observed data in each school to estimate the distribution of potential outcomes for all types of students ("common support"), or that extrapolation via the functional form of the model provides accurate estimates of the potential outcomes in those regions where there are no observed data ("functional form")

1.3 Estimating the Parameters of the Observed Data Model

School sample sizes are not under the control of the analyst. Particularly when interest focuses on value added within each grade for each school each year, the number of students available to estimate each value-added parameter may be rather small. The problem of small sample sizes is especially acute when some schools have small enrollment or when interest focuses on value added for subsets of students, e.g., ethnic minority students, students from low-income families, or students whose first language is not English. The prevalence of small sample sizes puts a premium on finding statistically efficient estimators. Efficiency is also helpful in coping with the inevitable problem of

missing data. Statistically efficient analysis of all available data weakens the required assumptions about the missing data process (Little & Rubin, 2002; Schafer, 1997). Sometimes efficient estimation requires computationally efficient algorithms (Lockwood et al., 2007).

In this paper, we shall not focus on statistical efficiency, missing data, or computational feasibility except to note that the press for efficiency pushes us in the direction of parametric rather than non-parametric models for the potential outcomes.

1.4 Organization of the Paper

In this paper, we have three aims:

1) to outline a potential outcomes counterfactual framework for understanding VAMs
2) to articulate the assumptions required for VAMs to provide unbiased, unambiguous rankings of the causal effects of teachers or schools on student achievement
3) to conduct a set of simulations to assess the sensitivity of value-added estimates to plausible violations of several of these assumptions.

We begin with a concise mathematical formulation of the general theoretical model for potential outcomes and causal effects, defining the need for (i) manipulability and (ii) no interference between units. We then add parametric assumptions, requiring (iii) an interval metric, and (iv) homogeneity of effects. Next, we consider the school assignment process, yielding the need for (v) strongly ignorable treatment assignment and (vi) functional form assumptions, insuring that the parameters estimable from the data are equivalent to the causal parameters of the theoretical model.

With these key assumptions in mind, we undertake a preliminary discussion of their plausibility and the likely departures from these assumptions in practice, with the aim of extending the discussion about the viability of the value added project. We then conduct a set of simple simulations to

investigate the effect of plausible departures from several of the assumptions. These simulations provide some guidance for future development of VAMs.

2. Definitions and Notation

First, some notation:

i indexes N students in population P

j indexes J schools[1] in population K

A is the *assignment matrix*, an $N\ x\ J$ matrix indicating the observed assignment of students $i = 1, \ldots N$ to schools $j = 1, \ldots, J$, where $a_{ij} = \mathbf{A}[i,j] = 1$ if $i \in j$, and $a_{ij} = \mathbf{A}[i,j] = 0$ if $i \notin j$.

θ_i is the true (unobservable) cognitive skill of student i at some time after school assignment (it is the true outcome measure). We cannot observe θ, but we assume it exists. In practice, because we cannot observe θ directly, we measure it using a cognitive test, yielding an observed test score Y_i. If the test is well-designed (meaning the test measures θ and no other skill or trait of the test-taker), then Y_i will be a function of θ plus some random measurement error: that is, $Y_i = g(\theta_i) + \varepsilon_i$, where *g* is a monotonically increasing function, and where ε_i is measurement error such that $E[\varepsilon_i|\theta] = E[\varepsilon_i] = 0$. Thus, we have

Y_{ij} is the (observed) achievement test score of student i in school j.

$\boldsymbol{x_i}$ is a vector of covariates for student i. In particular, $\boldsymbol{x_i}$ may contain θ_{i0}, the true cognitive skill of student i at time 0 (some point in time prior to assignment to a given school).

[1] To simplify discussion through the paper, we focus on the estimation of school, rather than teacher, effects. The basic logic of our argument would be unchanged if we considered the estimation of teacher effects instead. School value-added models are empirically somewhat simpler than teacher value added models, however, because they do not require one to consider the sorting of teachers among schools and the non-random sorting of students among teachers within schools.

2.1 Potential Outcomes Framework

In the Rubin model of causality, we define the average effect of treatment *T* versus treatment *C* in population *P* as the difference in average outcomes that members of *P* would experience if assigned to *T* rather than *C*. Conceptually, this requires that we believe that each member of *P* have two *potential outcomes*—the outcome they would experience if assigned to *T* and the outcome they would experience if assigned to *C*—despite the fact that they can experience only one of these. We adopt this potential outcomes framework to define precisely what we mean by school effects. We define the following:

Y_i^j is the *potential measured outcome* (the measured achievement test score) of student i (at some specified time) if the student were assigned to school j. Depending on the instrument used to measure θ, the measured outcomes will not be in the same metric as θ and may be measured with some error. That is, $y_i^j = y_i(a_{ij} = 1) = g\left(\theta_i^j\right) + \varepsilon_i$. Let $\mathbf{Y}$ denote the $N\ x\ J$ matrix of potential expected outcomes:

$$\mathbf{Y} = \begin{bmatrix} Y_1^1 & Y_1^2 & \dots & \dots & Y_1^J \\ Y_2^1 & Y_2^2 & \dots & \dots & Y_2^J \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ Y_N^1 & Y_N^2 & \dots & \dots & Y_N^J \end{bmatrix}$$

Each column of $\mathbf{Y}$ describes the vector of potential outcomes in the population if all students were assigned to a given school. We will use the following notation to summarize the distribution from which this vector of potential outcomes is drawn: let

G_P^j indicate the cumulative density function such that $G_P^j(y) = \text{Prob}\left(Y^j \leq y\right)$. That is, G_P^j is the distribution function governing the sample outcomes Y that would be realized if students in P

were assigned at random to school j.[2] To compare the effectiveness of two schools j and k, we compare the distributions G_P^j and G_P^k.

Formally, we define the average effect of attending school j rather than school k as δ_P^{jk} where $\delta_P^{jk} = r\left(G_P^j, G_P^k\right)$, and where r is some function that compares the distributions G_P^j and G_P^k (and where r is defined such that $r(a,b) = -r(b,a)$ and $r(a,b) > 0 \;\&\; r(b,c) > 0 \Rightarrow r(a,c) > 0$.)

Most commonly, for example, we define $\delta_P^{jk} = r\left(G_P^j, G_P^k\right) = \mu\left(G_P^j\right) - \mu\left(G_P^k\right)$, where $\mu\left(G_P^j\right)$ denotes the mean value of the distribution described by G_P^j. This yields an interval-scaled comparison of schools j and k. More generally, however, we can obtain an ordinal ranking of schools j and k by defining r as some function that returns a value of 1, 0, or -1 depending on whether school j ranks above, equal to, or below school k in effectiveness. For example, r might compare the medians (or other specified percentiles) of the distributions described by G_P^j and G_P^k. Alternately, we might define r such that $r = 1$ if G_P^j stochastically dominates G_P^k; $r = 0$ if $G_P^k = G_P^j$; $r = -1$ if G_P^k stochastically dominates G_P^j; and leave r undefined otherwise (if neither G_P^j nor G_P^k dominates the other).[3] The key point here is that a ranking of the effectiveness of two schools will depend on a comparison of their distributions of potential outcomes.

In general, then, in order to compare the effectiveness of two schools j and k we must estimate some parameter(s) of the distributions G_P^j and G_P^k. In order for this to be meaningful, it is necessary that Y_i^j and Y_i^k exist and are unique for all i. That is, it must be possible for each student to be assigned to

[2] If we assume there are no scale effects, then G_P^j describes the distribution of outcomes Y that would result if all students in P were assigned to school j.

[3] By stochastically dominates, we mean the usual: G_P^j stochastically dominates G_P^k if the cumulative distribution function of G_P^j is everywhere less than that of G_P^k.

schools j and k, and the potential outcome of student i if assigned to school j must not depend on what school other students attend. Formally, we require:

1. Y_i^j exists $\forall i \in P, j \in K$
2. $Y_i^j(\mathbf{A}) = Y_i^j(a_{ij}) \equiv Y_i^j$

Assumption 1 is the assumption of *manipulable treatment (school) assignment*. It states that it is meaningful to talk about the potential outcome of student i if she attends school j, because it is possible that student i could be assigned to school j. Following Holland (1986), we argue that is meaningless to talk about the relative effect of schools j and k for student i if there is no way that student i could attend one or the other of the two schools.[4] Consequently, it is meaningless to talk about the relative effectiveness of schools j and k in population P unless each student in P has non-zero probabilities (in principle) of attending both schools. We can only make causal statements about school effects for populations in which each student could attend any school.

Assumption 2 is the *stable unit treatment value assumption* (SUTVA) (Rubin, 1986). The SUTVA implies that that outcome of student i if assigned to school j does not depend the school assignment of any other student (that is, there are no peer effects). Under SUTVA, we treat each student as having J

[4] More specifically, school assignment must be *independently* manipulable—it must be possible that student i could attend both schools j and k while no other pre-exposure characteristics of i were changed. While it is possible, for example, that a high-income white student in Lake Forest (a wealthy, predominantly white suburb north of Chicago) could attend a low-income, all-black school in East St. Louis (at the southern end of Illinois), it is hard to envision how this could happen without some prior or concomitant change in his or her residential environment and family circumstances. If assignment to school j requires some change in other factors that might affect a student's potential outcome, then it is impossible—in principle—to identify the potential outcome Y_i^j resulting from assignment to j, and so it is impossible to define the estimand of interest, the causal effect of assignment to j rather than k, holding all else constant.

potential outcomes (one for each potential school assignment), of which we observe a single outcome. Without SUTVA, however, we must treat each student as having as many as J^N potential outcomes, one for each possible permutation of $\mathbf{A}$. Adopting the SUTVA may make the problem of causal inference tractable (but may do so at the cost of plausibility). We discuss later the plausibility of this assumption and the implications of its failure for VAMs.

Assumptions 1 and 2 are necessary to define the estimand of interest – the population average causal effect of school j relative to k – as we have done above. Given (1) and (2), comparing the effectiveness of two schools requires estimation of parameters of the distribution of potential outcomes for each school. Framing the value-added enterprise in this way distinguishes the dual issues involved in ranking schools: first, we must define what it means to say that the distribution of potential outcomes in one school is superior to another; second, we must estimate the distribution of potential outcomes for each school. That is, we must estimate the values in the matrix $\mathbf{Y}$ when we only observe one-J^{th} of the outcomes.

2.2 A Stylized Value-Added Model

In order to facilitate comparison of the distributions G_P^j and G_P^k, we typically write down a structural model such that $r\left(G_P^j, G_P^k\right)$ is recoverable from the parameters of the model. Most commonly, if r compares the means of G_P^j and G_P^k (that is, if $r\left(G_P^j, G_P^k\right) = \mu\left(G_P^j\right) - \mu\left(G_P^k\right)$), we write down a model that includes as parameters these means. For example, if the potential outcomes are described by a structural model of the form

$$Y_i^j = f(\boldsymbol{x_i}) + \Delta_j + \epsilon_i^j, \text{ where } \epsilon_i^j \perp x_i, j \;\&\; E\left[\epsilon_i^j\right] = 0 \qquad [1]$$

then the mean of $G_P^j = E\left[Y_i^j \middle| i \in P, j\right] = E[f(\boldsymbol{x_i})|i \in P] + \Delta_j$. The difference in the means of G_P^j and G_P^k is simply $\Delta_j - \Delta_k$. Under model [1], then, Δ_j and Δ_k identify the causal effect of attending school j relative to attending school k.

Embedded in this model, however, are two assumptions. First, the model requires that Δ_j is constant across students in P. This is the assumption of *school effect homogeneity*. Second, the model implicitly assumes that the metric of y_i^j is interval-scaled. This *interval-metric assumption* is necessary in order that a comparison of mean outcomes be a valid method of comparing the distribution of potential outcomes in two schools (the mean has no meaning in a non-interval-scaled metric). Moreover, the assumption of homogeneity is dependent on the metric—in general, if Y_i^* is some alternate scaling of θ (that is, $Y_i^* = g^*(\theta_i) + \varepsilon_i^*$, where g^* is a monotonically increasing non-linear function and ε_i^* is measurement error, then the school effect homogeneity assumption will not, in general, be valid under both g and g^*.[5]

To relax the homogeneity assumption, we can write the stylized model as

$$Y_i^j = f^j(\boldsymbol{x_i}) + \epsilon_i^j \qquad [2]$$

Model [2] replaces the function $f(\boldsymbol{x_i}) + \Delta_j$ in [1] with $f^j(\boldsymbol{x_i})$, allowing the difference in expected outcomes between schools to vary with $\boldsymbol{x}$. Under this model, the mean value $\mu\left(G_P^j\right) = E\left[f^j(\boldsymbol{x_i})\right] | i \in P$ and

$$\mu\left(G_P^j\right) - \mu\left(G_P^k\right) = \int_{\boldsymbol{x}}\left[f^j(\boldsymbol{x}) - f^k(\boldsymbol{x})\right]\rho(\boldsymbol{x})d\boldsymbol{x} \qquad [3]$$

where $\rho(\boldsymbol{x})$ is the density function of the vector $\boldsymbol{x}$ in the population P. Thus, to compare the average effectiveness of school j and k under this model requires estimation of f^j and f^k over the full range of $\boldsymbol{x}$ in P.

[5] To see this, assume homogeneity and assume no measurement error in either Y or Y^*. Now $Y^* = g^*\left(g^{-1}(Y)\right) = h(Y)$, where h is some nonlinear monotonic function. Homogeneity implies that $\bar{Y}^j | X = f(X) + \Delta^j$ (that is, the average potential outcome in school j is a function of the covariates X plus some school-specific effect Δ^j). Now $\bar{Y}^{*j} | X = h(Y) = h(f(X) + \Delta^j)$. Because h is non-linear, this cannot be separated into a function of X plus a function of j; the effect of school j in the transformed metric will depend on the value of X.

2.3 Estimating value-added models from observed data

In order to estimate the parameters of the distribution G_P^j we must make some additional assumptions, because the "fundamental problem of causal inference" ensures that we cannot observe the full distribution of potential outcomes for all schools (we observe at most only one-J^{th} of the potential outcomes in $\mathbf{Y}$). Estimation of the parameters of G_P^j and G_P^k from models of the form in [1] or [2] requires several additional assumptions.

First, we must assume *ignorability* (Holland, 1986): $(Y_i^j \perp A_{ij} | \boldsymbol{X_i} = \boldsymbol{x_i})$. Ignorability is a necessary, though not sufficient, condition for the observed data estimands of models like those in [1] and [2] to be equivalent to the causal parameters. In the absence of ignorability, the observed distribution of outcomes in school j, given $\boldsymbol{x}$, cannot provide an unbiased counterfactual estimate of the potential outcomes in j for students with $\boldsymbol{x}$ but not assigned to j.

Second, in order to estimate $f^j(\boldsymbol{x})$ for all values of $\boldsymbol{x}$, we must either observe a sample of cases where $a_{ij} = 1$ and $\boldsymbol{x_i} = \boldsymbol{x}$ or we must assume that we can extrapolate f^j from regions of $\boldsymbol{x}$ where there are observed cases with $a_{ij} = 1$ and $\boldsymbol{x_i} = \boldsymbol{x}$ into regions where there are no such observed cases (i.e., regions of $\boldsymbol{x}$ where $\rho(\boldsymbol{x}) > 0$ but $\rho^j(\boldsymbol{x}) = 0$. This is the *common support/functional form* assumption.[6]

[6] Rubin (1973) shows how a simple analysis of covariance can mislead when common support fails (that is, when some members of the treatment group have values on the covariate that are not shared by any members of the control group; and/or some members of the control group have values on the covariate that are not shared by any members of the treatment group). In these cases, the unmatchable cases, which by themselves carry no information about the treatment effect, can strongly influence the estimation of the regression coefficient for the covariate, leading to a poor statistical adjustment in the region of common support. Heckman, Ichimura, and Todd (1997) discuss bias arising from a failure of common support in the evaluation of job training programs using multivariate matching.

2.4 Are the Assumptions of VAMs Plausible? Manipulability, SUTVA and Ignorability

Above we have outlined six assumptions that underlie the definition and estimation of value-added models: 1) manipulability; 2) no interference between units; 3) interval-scale metric; 4) homogeneity of effect; 5) strongly ignorable treatment assignment; and 6) functional form. There are reasons to question the validity of each of these in some aspects of the school effects research. (There are other assumptions that may be required for the unbiased estimation of school effects and their standard errors—such as the absence of measurement error in **x**, normality and homoskedasticity of errors, and the independence of observations within schools—but we will not dwell on these here).

Manipulability. As we note above, manipulability requires that it be possible or at least theoretically interesting to conceive of each student attending any school in the population without necessarily altering any other pre-enrollment characteristic or condition of the student. Without manipulability, some potential outcomes—and some estimands, therefore—are not defined. It certainly stretches the bounds of common sense to think that any student could attend any school with no other pre-enrollment change in his or her conditions, though whether we wish to call it impossible may be debatable. Rather than consider the somewhat philosophical argument about whether it all potential outcomes exist in theory, for the remainder of this paper, we assume manipulability for the sake of argument. However, we note that the near-zero likelihood of observing some potential outcomes in practice suggests the inappropriateness of focusing on ranking all schools in a state, for example. If there are large populations of students who, in all likelihood, would never attend a given school, it is of little policy relevance to estimate the effect of that school on them. Moreover, as we note below, this practical reality is also likely to substantially limit common support in the observed data.

No Interference Between Units. Recall that SUTVA requires that the potential outcomes of one student

must not depend on the school assignments of other students. Strictly speaking, this means that a given student's achievement gains in a particular school do not depend on who his schoolmates are (or even how many of them there are). The literature on peer effects and organizational theory suggests this may be an implausible assumption (Angrist & Lang, 2004; Barr & Dreeben, 1983; Boozer & Cacciola, 2001; Hoxby & Weingarth, 2006). If student composition affects instructional practices and curricula, for example, and if what and how teachers teach affects student learning, then SUTVA will be violated. Likewise, if student composition affects the ability of a school to attract and retain quality teachers, then SUTVA will be violated. The consequences of SUTVA violations on the estimates of school effects are unclear, since without SUTVA the estimands of interest are not well-defined.

Ignorability. In order that parameter estimates obtained from models [1] or [2] be unbiased, we require school assignment to be *ignorable*, conditional on the observed covariates $\boldsymbol{x}$. To say that school assignment is ignorable within levels of covariates $\boldsymbol{x}$ states that for persons who share the same values of the covariates $\boldsymbol{x}$, the assignment to schools is independent of the potential outcomes of those students (that is, it is as good as randomly assigned). Formally, we write $A_{ij} \perp Y_i(a_{ij})|\boldsymbol{X} = \boldsymbol{x}$. This means that once we control for the covariates, a student's assignment to a given school does not depend on how well that student would do if assigned to that school. Under this assumption, the mean observed outcome of those students who are actually assigned to a given school is the same as the population mean of the potential outcomes of all students if assigned to that school. Formally, $E\left[Y_i^j \middle| A_{ij} = 1, \boldsymbol{X} = \boldsymbol{x}\right] = E\left[Y_i(a_{ij} = 1) \middle| A_{ij} = 1, \boldsymbol{X} = \boldsymbol{x}\right] = E\left[Y_i(a_{ij} = 1) \middle| \boldsymbol{X} = \boldsymbol{x}\right]$ (within levels of $\boldsymbol{x}$, the average value of the observed outcomes Y_i^j for students assigned to school j will equal the average of the potential outcomes $Y_i(a_{ij} = 1)$ for all students sharing the same values of $\boldsymbol{x}$).

Value-added models lean heavily on the ignorability assumption (as do most non-experimental studies). In practice, researchers assume that models that ignorability is plausible in models that

condition on a large vector of prior student achievement tests and/or student fixed effects. The consequence of non-ignorable school assignment is potential bias in value-added estimates. A violation of the ignorability assumption would imply that the outcomes we observe among students in a given school do not provide an unbiased estimate of the average outcomes we would observe if all students were assigned to a given school, even after we adjusted our estimates using observed covariates. In other words, absent ignorability, we cannot distinguish a good school (one that produces larger than average achievement) from an average quality school that selectively enrolls students with better than expected outcomes. In general, however, the ignorability assumption is unverifiable unless the assignment mechanism is known (and observed), such as in random assignment or some other observable assignment mechanism.

Although the ignorability assumption cannot be formally proven, it does have testable implications that may allow researchers to assess its plausibility in a given sample. Rothstein (2007), for example, conducted a falsification test and found that teacher value-added models provide estimates of teacher effects on prior achievement that are nearly as large as the estimates of teacher effects on current achievement, strongly suggesting that teacher assignments are not ignorable, even within schools. Although we are not aware of similar research that has conducted falsification tests of the ignorability of school assignment, the Rothstein result should make researchers cautious of assuming ignorability in teacher or school effects research. Given that sorting among schools is generally more pronounced than among classrooms within schools, it is likely that some aspects of sorting among schools are related to students' potential outcomes in ways not captured by observable student characteristics.

2.5 Plausibility of Assumptions: Homogeneity, Functional form, and Interval scale

As noted above, common sense, theory, and empirical evidence each provide reasons to suspect

that the assumptions of manipulability, SUTVA, and ignorability do not hold in many cases where we may wish to estimate value-added models. Violations of these assumptions render value-added models either intractable (as in the case of violations of manipulability and SUTVA) or biased (in the case of violations of ignorability). We now turn our attention to the consequences of violations of the remaining three assumptions: homogeneity, functional form/common support, and interval scaling. Our goal here is to focus on these three relatively overlooked assumptions in order to assess their potential consequences, which are less well understood than the first three. To do so, we will assume that manipulability, SUTVA, and ignorability hold, despite our suspicions that they do not, in order to isolate the consequences of the remaining three assumptions. We begin by considering the plausibility of the homogeneity, functional form/common support, and interval metric assumptions. Following this, we conduct a set of simulation analyses to investigate how violations of these assumptions affect the parameter estimates from value-added models.

Homogeneity. The strictest form of homogeneity—requiring that a school has the same effect on every student who attends—is not credible. Nonetheless, the homogeneity assumption implied in model [1] above requires only a weaker form of homogeneity, the assumption that the difference in mean potential outcomes is the same regardless of student covariates $\boldsymbol{x}$. Weaker still is a version of this assumption that we might term monotonicity – the assumption that if a school is better for one subset of students (defined by $\boldsymbol{x}$), it is better for all other subsets. Although typical value-added models assume some form of homogeneity, there are many reasons to think this assumption is not strictly true. If schools target their curricula to students in the middle of the school's skill distribution, then schools' relative effectiveness may vary across students' prior skill levels. Schools where English Learner students are a majority may be more effective for such students than schools where such students are a small minority, because the former schools may have more specialized curricula and instruction for

English Learners. In our simulations, we assess the extent to which violations of homogeneity lead to incorrect rankings of schools.

Functional form/common support. The common support assumption is clearly invalid to some extent in practical applications of value-added models. Research on school effects finds that 15-25% of the variance in observed test scores lies between schools, implying unequal distributions of achievement levels among schools (Hedges & Hedberg, 2007). In the presence of such sorting, it is clear that not all schools will have students at all skill levels (or not sufficient numbers at all skill levels to provide precise estimates of mean achievement gains at each skill level). In the absence of common support, value-added models rely heavily on the assumption that their functional form is correct and allows valid extrapolation into regions with no observed data. In fact, the homogeneity assumption is, strictly speaking, a functional form assumption, which allows estimation of the potential outcomes in a given school for types of students who are not found in that school.

In the case of a simple value-added model such as that in [3] above, estimation of school effects requires extrapolation of the functions f^j and f^k to values of $\boldsymbol{x}$ where there are no students with $\boldsymbol{x}$ in j and/or k. The most common functional forms in education are polynomial forms. Such forms are effective in providing local approximations of a wide variety of continuous functions. So if matched pairs $\boldsymbol{x}, y$ have been observed over a restricted domain of $\boldsymbol{x}$, a polynomial function $Y \approx f(\boldsymbol{x})$ may provide a reasonable approximation. However, polynomials may be grossly inaccurate at extrapolation beyond the domain of *x* that supplied the local data.

The question of interest for our simulation is whether extrapolations based on polynomial models are justifiable in light of the degree to which U.S. schools are segregated on background variables $\boldsymbol{x}$ that predict student achievement Y. We use extant data on such segregation in the U.S. to inform our choice of parameters in the simulation study.

Interval scale metric assumption. The choice to compare distributions of potential outcomes using a comparison of the means (i.e., to rely on $r\left(G_P^j, G_P^k\right) = \mu\left(G_P^j\right) - \mu\left(G_P^k\right)$ to rank schools) implies that we are treating Y as interval-scaled. To see this, consider a distribution G_P^j. Suppose the distribution of potential outcomes in school k, G_P^k, is identical to G_P^j except that the potential outcome of one student h is lower by an amount c if assigned to k than to j. That is, suppose the potential outcomes in school k are given by

$$Y_i^k = \begin{cases} Y_i^j & \text{if } i \neq h \\ Y_i^j - c & \text{if } i = h \end{cases}$$

Now, $r\left(G_P^j, G_P^k\right) = \mu\left(G_P^j\right) - \mu\left(G_P^k\right) = \frac{c}{N}$ (where N is the size of the population), regardless of whether student h is a student whose potential outcome in school j, Y_h^j, would have been high or low. The function r in this case implicitly values a one unit difference in scores equally at all points in the range of Y. In other words, if we compare the distributions of potential outcomes in schools by taking the difference of their means, then we have implicitly treated Y as if it is interval-scaled. Likewise, if we consider Y to be interval-scaled, then a comparison of means is sufficient to determine which of two distributions contains the greater total quantity measured by Y. Thus, regression models that estimate conditional means implicitly assume an interval-scaled metric. If the metric of Y_i is a non-linear transformation of some true, interval-scaled metric, then ranking schools by their mean potential outcome may lead to different relative ranks than would obtain from a comparison of means of the original interval metric.

It is unclear how one can determine whether a given test metric should be considered interval-scaled. To be interval-scaled is to be linearly-related to some reference metric. Distance, mass, temperature, and time are defined in interval metrics via their correspondence with tangible physical quantities, for example (e.g., a meter is the distance that light travels in a vacuum in 1/299,792,458th of

a second; this definition is independent of place and time, so a meter has the same meaning, in a tangible sense, regardless of where we are). In the case of cognitive skill, however, it is unclear what the reference metric should (or could) be. We cannot observe cognitive skill except by inferring it from some behavioral task (such as a test), to which we must then assign a metric, but we cannot know if that metric is itself interval-scaled. As such, there is often no clear reference metric for cognitive skill.[7] In the absence of such a reference metric, we could anchor it to some other correlate of skill (average time required to gain a given level of skill; average value of a difference in skill in the marketplace, etc), but any such anchor will likely be variable and subject to changing social conditions (the average time it takes to gain a given level of skill depends on schooling practices, which may change over time; the average income return to differences in skill depends on the supply and demand for given skills in the economy, which also clearly change over time). This contextual dependence of such approaches render them unsatisfactory as methods of defining an interval metric.

In cases where it is impossible to claim with certainty that a given test score metric is interval-scaled, it would be useful to know the extent to which inferences from value-added models are sensitive to monotonic transformations of test scores. Our simulations assess the sensitivity of simple value-added estimates to modest violations of the interval scaling assumption.

[7] An argument may be made that a Rasch scale is in fact an interval scale—in a Rasch scale, a given difference in ability corresponds to a constant difference in the log-odds of answering a given item correctly, regardless of the difficulty of the item or the ability of the test-taker. In a real sense, then, the Rasch scale uses as a reference metric performance on a behavioral task in a well-defined sense. Under the assumptions of the Rasch model (unidimensionality of the domain, the log-odds functional form of the item-response curve), the Rasch model yields an interval metric. Ballou (2008), however, argues that, in practice, achievement tests often do not meet the conditions required to satisfy the assumptions of the Rasch model.

3. Simulation Analyses

In order to investigate the effect of violations of the homogeneity, functional form/common support, and metric assumptions, we specify a (simple) 'true' potential outcome generating model describing the outcomes for each student i if assigned to each school j (that is, we specify parameters describing the full potential outcomes matrix $\boldsymbol{Y}$). From this model, it is simple to recover the 'true' school effects for each school. We then define the parameters describing the observed data (that is, we specify parameters describing which potential outcomes we observe). Next, we simulate data according to these parameters. We then fit several (very simple) value-added models using these data, and use their estimated parameters to recover estimates of the school effects for each school. Finally, we compute the correlations among the estimated and `true' school effects. By varying key parameters describing the 'true' potential outcomes and the observed data, we can assess the sensitivity of our conclusions regarding school effects to violations of the VAM assumptions. We use large enough sample sizes in the simulations to insure that inaccuracies in estimation largely reflect failures of model identification while sampling error plays a comparatively trivial role.

A model for the true potential outcomes. We assume that the true potential outcomes for student i are given by a simple hierarchical linear model (Raudenbush and Bryk, 2002). Under this model, there is a single relevant covariate, namely the pre-test Y_{i0}. Within each school, the observed outcome is a linear function of this covariate plus random noise. The average effect of attending the school is the school-specific random intercept of this model, while heterogeneity of the school effects is captured by the randomly varying coefficient of the pretest Y_{i0}. This model provides a comparatively simple framework for examining the plausibility of key assumptions that underlie value added analyses. We do not mean to imply, of course, that the simple hierarchical linear model is an adequate description of the value added process in the real world or that fitting of such models is an optimal strategy for value added

analysis. Rather, this simple simulation exercise is meant to help us understand the extent to which value-added model estimates are sensitive to a set of assumptions.

Formally, the model generating our data is given by the school-specific model

$$Y_i^j = \beta_{0j} + \beta_{1j}\left(Y_{i0} - \bar{\bar{Y}}_0\right) + e_{ij}$$

combined with the between-school equations

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

where

$$e_{ij} \sim N(0, \sigma^2)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix}\right]. \quad [4]$$

We specify the parameters of the model that describes the true potential outcomes as in [4] above. The key parameter of interest with respect to the homogeneity assumption is τ_{11}, which defines the variance of β_{1j} (the association between Y_{i0} and the potential outcome) across schools. If $\tau_{11} > 0$, then we have heterogeneity of school effects with respect to Y_0. In order to assess the sensitivity of value-added rankings to heterogeneity of school effects, we will vary τ_{11}.

Under this model, it is simple to show that the average potential outcome if all students are assigned to school j is given by

$$\mu\left(G_P^j\right) = \gamma_{00} + u_{0j} \quad [5]$$

Thus, the effect of school j relative to the average school in the population is determined by u_{0j} (since γ_{00} is constant—it represents the average potential outcome in the average school in the population—and without loss of generality we can set it to 0). Note that this will true even under a model that includes other covariates $\boldsymbol{x}$ (as long as x is centered on its mean), though we do not include other

covariates here for simplicity.[8]

Simulating the observed data. Once we have specified a set of parameters describing the potential outcomes model, we can compute the true effect of each school from [5] above. Next, we simulate a sample of observed data. For this, we must specify how students are sorted among schools with respect to Y_{i0}, u_{0j} and u_{1j}. The key parameter of interest here is the variance of the school mean values of y_{i0}, $\omega = Var(\bar{Y}_{j0})$ (we set the $Var(Y_{i0}) = 1$ for simplicity, so ω is the proportion of variance in Y_{i0} that lies between schools). If $\omega = 0$, there is no sorting of students among schools in relation to their initial scores, implying full common support in the data; $\omega > 0$ implies sorting and weakened common support.

Generating Transformations of *Y*. We are interested in assessing the sensitivity of our rankings to nonlinear monotonic (increasing) transformations $g(Y_i)$. For simplicity, we let g be a quadratic function that is monotonically increasing over the observed domain of the outcome Y_i. We define the curvature κ_g of g over the range of Y_i to be the ratio of the slope of $g(Y_i)$ at the 95th percentile of Y_i to the slope at the 5th percentile of Y_i. This provides a simple measure of the extent of nonlinearity in g over the domain of Y_i. $\kappa_g > 1$ implies a transformation of the metric that exaggerates score differences at the higher end of the distribution; $0 < \kappa_g < 1$ implies a transformation of the metric that exaggerates score

[8] If the potential outcomes are given by a model that is non-linear in Y_{i0}, such as

$$Y_i^j = \beta_{0j} + \beta_{1j}\left(Y_{i0} - \bar{\bar{Y}}_0\right) + \beta_{2j}\left(Y_{i0} - \bar{\bar{Y}}_0\right)^2 + e_{ij}$$

then the mean potential outcome under assignment to j will not be given solely by $\gamma_{00} + u_{0j}$. Rather, it will be:

$$\begin{aligned}\mu\left(G_P^j\right) &= \beta_{0j} + \beta_{2j}Var(Y_{i0}) \\ &= \left(\gamma_{00} + \gamma_{20}Var(Y_{i0})\right) + \left(u_{0j} + u_{2j}Var(Y_{i0})\right)\end{aligned}$$

In this case, the effect of school j is determined by u_{0j} and u_{2j} (since γ_{00}, γ_{20}, and $Var(Y_{i0})$ are constant).

differences at the lower end of the distribution ($\kappa_g = 1$ implies a linear transformation).

Simulation model comparisons. Under our model for the potential outcomes [4], the 'true' effect of school j relative to the average school in the population is simply $\mu\left(G_i^j\right) = u_{0j}$. We specify four alternate models for estimating the school mean potential outcome $\mu\left(G_i^j\right)$ (the school effect) from the simulated observed data.

First we fit model A, a model that assumes homogeneity of school effects and a linear relationship between Y_i and Y_{i0}:

$$Y_{ij} = \gamma_{10}^A\left(Y_{i0} - \bar{\bar{Y}}_0\right) + \Delta_j^A + e_{ij}^A \qquad [6]$$

The school effect $\mu\left(G_i^j\right)$ in this model is simply $\widehat{\Delta}_j^A$.

Second we fit model B, a model that assumes homogeneity of school effects, but that allows the relationship between Y_i and Y_{i0} to be nonlinear (we allow a quadratic specification). Even if the true relationship between Y_i and Y_{i0} is linear (as specified in [4]), a nonlinear transformation of Y_i will render the observed relationship nonlinear, potentially confounding estimates from model A even if the homogeneity assumption in accurate. Model B allows the model to fit the nonlinearity more appropriately, so that we can assess the extent to which the homogeneity assumption in the model confounds estimates of school effects when the data generating process includes heterogeneity. Model B is:

$$Y_{ij} = \gamma_{10}^B\left(Y_{i0} - \bar{\bar{Y}}_0\right) + \gamma_{10}^B\left(Y_{i0} - \bar{\bar{Y}}_0\right)^2 + \Delta_j^B + e_{ij}^B \qquad [7]$$

The school effect $\mu\left(G_i^j\right)$ in this model is simply $\widehat{\Delta}_j^B$.

Models A and B assume homogeneity, as do typical value-added models. In models C and D, we allow for heterogeneity of school effects, by allowing the association between Y_i and Y_{i0} to vary among schools, by adding random effects to models A and B. Model C is:

$$Y_{ij} = \left(\gamma_{10}^{C} + u_{1j}^{C}\right)\left(Y_{i0} - \bar{\bar{Y}}_0\right) + \Delta_j^C + e_{ij}^C \quad [8]$$

The school effect $\mu\left(G_i^j\right)$ in this model is $\widehat{\Delta}_j^C$. Because model C is identical to the data generating model, and because we have assumed ignorable school assignment, conditional on Y_{i0}, $\widehat{\Delta}_j^C$ will be an unbiased estimate of the 'true' school effect if the outcome Y_i is measured in the same metric as the original data (if $\kappa_g = 1$). If, however, Y_i is measured in a transformed metric (if $\kappa_g \neq 1$), then $\widehat{\Delta}_j^C$ may be biased.

Model D allows both the linear and quadratic terms in model B to vary across schools, and so may be less sensitive to metric transformations than is B or C. Specifically, model D is:

$$Y_{ij} = \left(\gamma_{10}^{D} + u_{1j}^{D}\right)\left(Y_{i0} - \bar{\bar{Y}}_0\right) + \left(\gamma_{20}^{D} + u_{2j}^{D}\right)\left(Y_{i0} - \bar{\bar{Y}}_0\right)^2 + \hat{\delta}_j^D + e_{ij}^D \quad [9]$$

The school effect $\mu\left(G_i^j\right)$ in this model is $\widehat{\Delta}_j^D = \hat{\delta}_j^D + \hat{u}_{2j}^D Var(Y_{i0})$. Note that models C and D allow for heterogeneity of school effects across Y_{i0}. In this regard, they are not typical of value-added models used in practice.

We fit models A and B using a fixed-effects estimator, and fit models C and D using a random-coefficients estimator (Raudenbush & Bryk, 2002). Because we simulate data with large numbers of students within schools ($n = 500$), and because we assume no measurement error in Y, we obtain highly reliable estimates of the $\widehat{\Delta}_j$'s.

Models A-D provide four estimates of $\mu\left(G_i^j\right)$. We next examine the correlations between the true value of u_{0j} and each of the estimated values of $\mu\left(G_i^j\right)$: $\widehat{\Delta}_j^A$, $\widehat{\Delta}_j^B$, $\widehat{\Delta}_j^C$, and $\widehat{\Delta}_j^D$. These will be, in a number of regards, best-case correlations, for a number of reasons: 1) we have assumed a data generating model in which the assumption of ignorability is valid; 2) we assume no measurement error in either Y_{i0} or Y_i; 3) we simulate data with very large within-school samples, so the effect of sampling variance is trivial; 4) we vary only the metric of the outcome Y_i, not the metric of Y_{i0}; 5) we apply only simple quadratic transformations to Y_i; and 6) we assume a very simple (linear in Y_{i0}) data generating process.

Parameters used in simulations. There are three parameters of interest in our simulations: τ_{11}, which specifies the degree to which school effects are heterogeneous with respect to Y_{i0}; ω, which specifies the extent of common support across schools; and κ_g, which indicates the extent of non-linearity in the outcome metric. We choose values of $\tau_{11} = \{0, .2, .4\}$ (ranging from 0—which is what is assumed by homogeneity—to .4, which is a plausibly high value) for the heterogeneity parameter. For the intracluster correlation, we use values of $\omega = \{.1,\ .2,\ .3\}$, based on the empirical studies of intracluster correlations of test scores (Hedges & Hedberg, 2007). For the curvature parameter, we use values of $\kappa_g = \left\{\frac{1}{5}, \frac{1}{3}, \frac{2}{3}, 1, 1.5, 3, 5\right\}$. These are within the range of plausible values implied by empirical studies; for example, Koedel and Betts (2007) show data where average reading gain scores for students with initial scores in the first decile are three times larger than those of students with initial scores in the tenth decile, implying a curvature parameter of 1/3 relative to a test where average gain scores are equal across the range of initial scores. We fix all other parameters in the simulation.[9] We simulate random samples of 500 schools, each containing 500 students.

3.1 Simulation Results

Figures 1-4 report the results of the simulation exercise described above. Figure 1 displays the correlations of the estimated school effects from model A and the true school effects. Recall that model A assumes homogeneity of school effects and a linear association between initial score and the outcome score. As expected, when the true potential outcomes model includes no heterogeneity of school

[9] We set $\tau_{00} = .4$ (implying an intracluster correlation of gain scores of .4); $\tau_{01} = 0$; $Corr\left(\bar{Y}_{j0}, u_{0j}\right) = .25$ (implying a positive correlation between students with higher initial scores are disproportionately assigned to schools that have larger effects for average skill students); $Corr\left(\bar{Y}_{j0}, u_{1j}\right) = .25$ (implying a positive correlation between students with higher initial scores are disproportionately assigned to schools that have larger effects for high initial skill students than low initial skill students).

effects ($\tau_{11} = 0$) and when the metric has little or no curvature, model A produces estimates that are nearly perfectly correlated with the true school effects, regardless of the amount of sorting (ω) of students among schools on their initial scores. In other words, the fact that students are unevenly distributed among schools with respect to their prior achievement (resulting in a lack of common support to estimate school effects) need not produce error in the estimated school effects, so long as the true school effects are homogeneous. The assumption of homogeneity is a strong functional form assumption; if it is valid, the estimation of school effects can be accomplished even in the absence of common support.

In the presence of substantial curvature, however, the correlations are less than perfect, albeit quite high. Likewise, when there is heterogeneity of school effects ($\tau_{11} > 0$) model A is less successful at reproducing the true school effects, even in the absence of curvature of the observed test score metric relative to the true metric. The consequences of violations of both the homogeneity and metric assumptions are exacerbated when there is considerable sorting of students among schools.

Figure 2 reports the corresponding correlations estimated from model B. Like model A, model B assumes homogeneity of school effects, but allows the association between initial score and observed achievement to be nonlinear (quadratic in this case). While we might not expect this to reduce the errors in the estimates that results from heterogeneity, it may make the estimates less sensitive to the curvature of the metric, since the nonlinearity in the model may be able to account for some of the nonlinearity in mean outcomes induced by the transformation of the test metric. The results in figure B generally support this hypothesis. In general, the correlations in figure B are very similar to those in figure A, particularly when the curvature parameter is near 1. Using a more flexible functional form does nothing to alleviate the misestimation of school effects caused by the interaction of heterogeneity of effects and sorting among schools on initial score. When the curvature is large, however, the flexible functional form of model B produces estimates slightly more similar to the true effects than does model

A, but the difference is relatively modest.

As we might expect, school effects are heterogeneous, then a model that explicitly allows for such heterogeneity is much more successful at producing accurate rankings of schools. Figures 3 and 4 report the results of models C and D, both of which allow for heterogeneous school effects. Model D differs from C in that it allows the association between initial score and observed achievement to be nonlinear (quadratic in this case). Both models C and D produce dramatically better estimates than do models A and B. When the curvature parameter is near 1, models C and D yield estimates virtually perfectly correlated with the true school effects, regardless of the degree of heterogeneity of the effects or the sorting of students among schools by initial score. The estimates are somewhat sensitive to transformations of the outcome test metric, and this sensitivity is greater as the heterogeneity of effects grows larger. In general, the results of models C and D suggest the importance of modeling the heterogeneity of school effects in value-added models absent a strong reason to believe the effects are truly homogeneous.

Our simulations provide information regarding the conditions under which violations of the common support, homogeneity, and metric assumptions lead to invalid inferences regarding the ranking of schools in a value-added model framework. First, note that the fact that students are assigned to schools based on prior achievement does not in itself invalidate inferences. This is a case of sorting based on an observed covariate, a kind of sorting that is acceptable in the framework of ignorability conditional on *x*. However, if the school effect depends on the covariate, and if this heterogeneity is not specified in the analytic model, the estimated school effects degrade, as is evident in the case of Models A and B. This tendency of sorting to degrade accuracy appears a bit more pronounced when the assumption of linearity fails. Put another way, the failure of linearity has its most pronounced effects on the estimates in the presence of heterogeneity—unless the heterogeneity is represented in the model for the observed data.

4. Conclusion

In this paper we have applied the counterfactual model of causality to the problem of value-added modeling. This analysis reveals six assumptions that are needed to make valid causal inferences about school effects on student learning. Three of these assumptions concern our conception of the causal effects themselves even in the absence of any attempt to assess them. First, we assume *manipulability*, that is, every student is potentially assignable to every school. Given the segregation of U.S. schools, this assumption is not plausible, but it is hard to conceive of the value added project going forward without it. Second, we assume no interference between children, that is, that the impact of attending a school does not depend on the school assignments of other children. This assumption is also implausible in that the composition of a school, particularly on the basis of students' prior cognitive skill, sets conditions for the design and enactment of classroom instruction. Hong and Raudenbush (2006) extend the potential outcomes framework to allow for such compositional effects in the case of studying explicit school policies, but doing so in the context of value-added model framework, while possible, is considerably more challenging. The third assumption is one of homogeneity – a school that is better for one sub-population of students is better for every other sub-population. This is certainly a contestable assumption.

The other three assumptions arise in designing a study to estimate the school effects. First, we must assume that the assignment of children to schools does not depend on their potential outcomes once we have controlled for observed covariates. One can argue that such an assumption is not entirely implausible if one has access to highly valid and reliable measures of prior cognitive skill, though this assertion can readily be critiqued. The remaining two assumptions are related: we typically want to represent the achievement process as some kind of continuous function of prior background and school impact. The specification of functional form and the metric are tightly connected. These are linked also to the problem of lacking "common support." In this case, the data suggest that certain sub-sets of

students have little or no risk of attending certain schools. Such a problem may emanate from a failure of the manipulability assumption or may simply reflect the fact that the sample data are too sparse to reflect a population in which all students have at least a small probability of attending any school. The lack of common support requires the value-added modeler to use assumptions about the functional form to extrapolate expected effects of attending each school for subsets of kids with little risk of attending certain schools.

In our simulation, we stipulated the assumptions of manipulability, no interference between units, and ignorable school assignment so that we could focus on the interplay between homogeneity, functional form, and common support. While our simulations were relatively simple and somewhat artificial, they do point to some important considerations in implementing value-added models. Perhaps the most significant result involves the importance of modeling heterogeneity when it exists. Failure to do so increases the negative effects of sorting and curvature. Explicitly modeling heterogeneity protects against these negative effects. This is a significant finding for practice because value-added systems rarely estimate heterogeneous effects of schools.

It is essential to emphasize that our simulation has accepted as fact three key assumptions that are implausible and the violation of which may substantially degrade results. The extent to which such violations are influential is a crucial topic for future research. Moreover, our data generating model is intentionally made simple to illustrate basic principles. For that reason, it is easy to "get right." More realistically complex models for the true effects may pose significant analytic challenges. Finally, we have simulated large samples in order to insure that errors of inference primarily reflect failures of model identification and not estimation error. In practice, when sample sizes are more modest, failure of identification and sampling error will conspire to reduce correlations between estimated and true school effects.

References

Angrist, J., & Lang, K. (2004). Does school integration generate peer effect? Evidence from Boston's Metco Program. *American Economic Review, 94*(5), 1613-1634.

Ballou, D. (2008). *Test Scaling and Value-Added Measurement*. Paper presented at the National Conference on Value-Added Modeling.

Barr, R., & Dreeben, R. (1983). *How Schools Work*. Chicago, IL: University of Chicago Press.

Barr, R., & Dreeben, R. (1988). Classroom Composition and the Design of Instruction. *Sociology of Education, 61*(129-142).

Boozer, M., & Cacciola, S. (2001). *Inside the Black Box of Project STAR: Estimation of Peer Effects using Experimental Data*: Yale University.

Cox, D. R. (1958). *The Planning of Experiments*. New York: Wiley.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Heckman, J. J. (1979). Sample selection bias as specification error. *Econometrica, 47*(1), 153-161.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies, 64*, 605-654.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945-970.

Hong, G., & Raudenbush, S. W. (2006). Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference For Multi-Level Observational Data. *Journal of the American Statistical Association, 101*(475), 901-910.

Hoxby, C. M., & Weingarth, G. (2006). Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects. *Unpublished manuscript*.

Koedel, C., & Betts, J. R. (2007). *Re-Examining the Role of Teacher Quality in the Educational Production Function* (No. 07-08): University of Missouri, Department of Economics.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). New York: John Wiley and Sons.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setoldi, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125-150.

Neyman, J. S. (1990). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9. *Statistical Science, 9*(4), 465-472.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

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

Rothstein, J. M. (2007). Do Value-Added Models Add Value? Tracking, Fixed Effects, and Causal Inference.

Rubin, D. B. (1973). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics, 29*(1), 184-203.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics, 6*, 34-58.

Rubin, D. B. (1986). Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association, 81*(396), 961-962.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Boca-Raton, FL: Chapman-Hall.

Figure 1

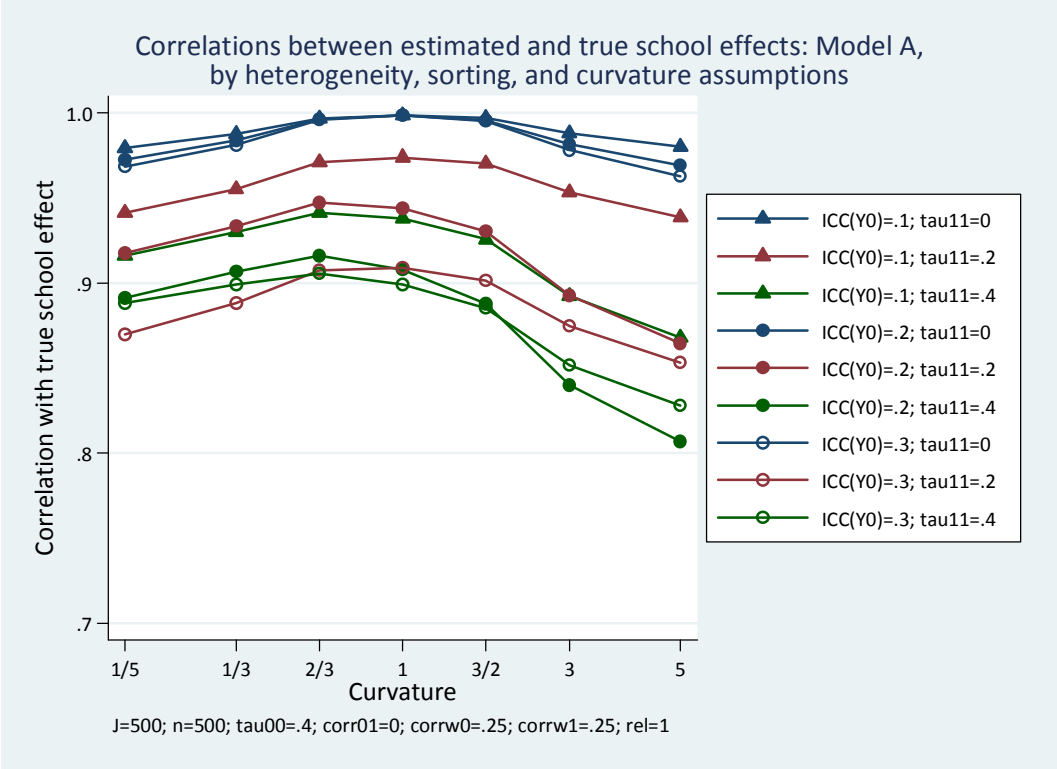

Figure 2

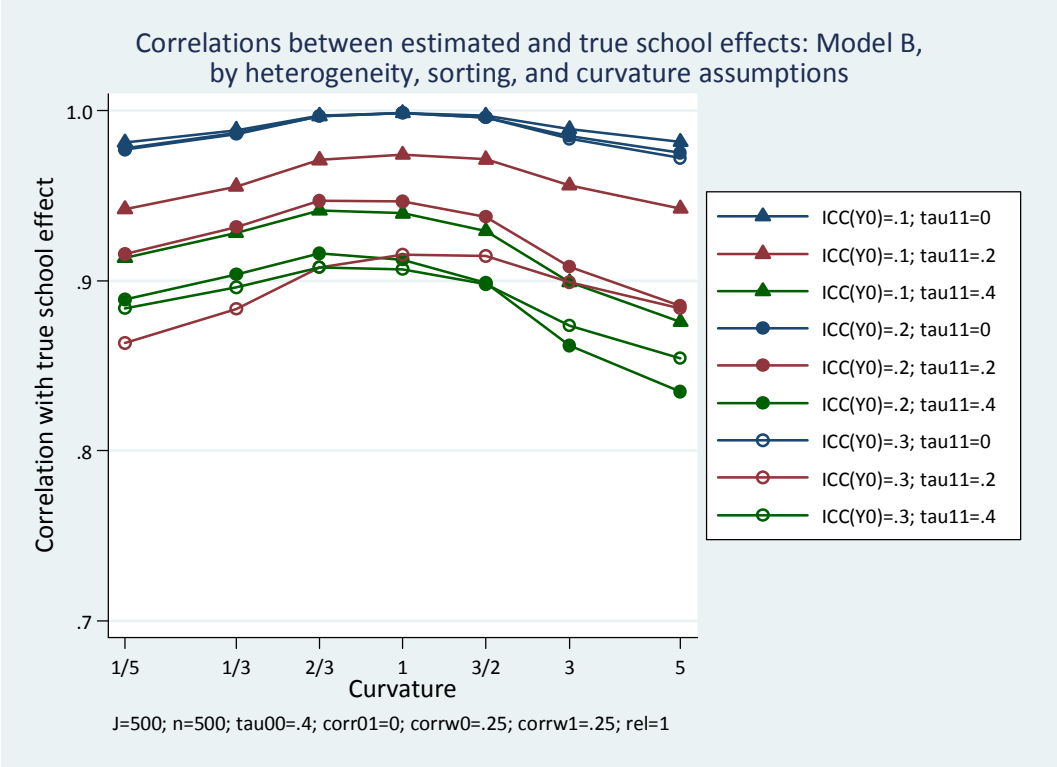

Figure 3

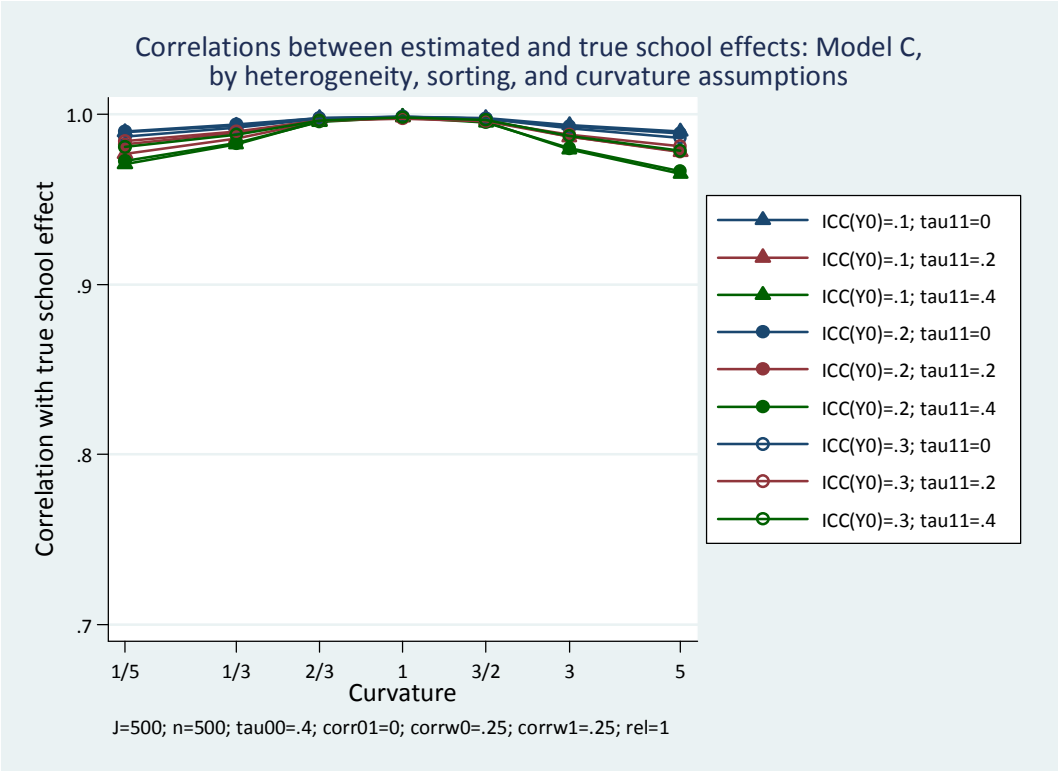
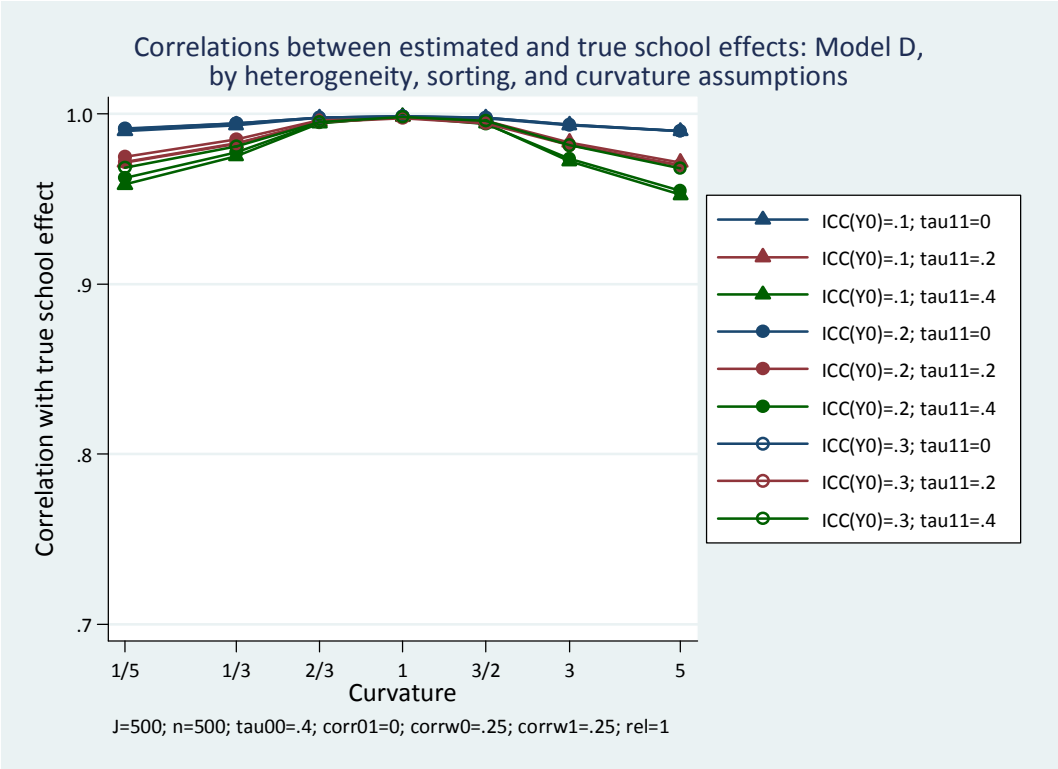
Figure 4