Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2018, Vol. 37, No. 2, pp. 45–53

Examining Estimates of Intervention Effectiveness Using Sensitivity Analysis

Chen An, *Department of Research & Evaluation, Orange County Public Schools, Orlando, FL* Henry Braun and Mary E. Walsh, *Boston College, Chestnut Hill, MA*

Making causal inferences from a quasi-experiment is difficult. Sensitivity analysis approaches to address hidden selection bias thus have gained popularity. This study serves as an introduction to a simple but practical form of sensitivity analysis using Monte Carlo simulation procedures. We examine estimated treatment effects for a school-based support intervention designed to address student strengths and needs in academic and nonacademic areas by leveraging partnerships with community agencies. Middle school (Grades 6–8) statewide standardized test scores in mathematics and English language arts (ELA) were examined for students in a large urban district who participated in City Connects during elementary school. Results showed that the estimated treatment effects in both subjects were reduced slightly with the inclusion of U, a hypothesized unobserved binary variable. However, simulated effects fell within one-sided 90% confidence intervals for original treatment effects, suggesting only a mild sensitivity to hidden bias. Moreover, almost identical estimated treatment effects were observed when the magnitude of the mathematical difference between each pair of the conditional probabilities of U given the treatment indicator Z was the same.

Keywords: causal inferences, Monte Carlo simulation, program effectiveness, sensitivity analysis

This study performed a sensitivity analysis of the estimated effectiveness of a kindergarten through fifth grade student support intervention on academic achievement during middle school. The study serves two purposes: (1) to examine to what extent the estimated treatment effects are robust to the presence of unobserved selection bias that may jeopardize causal inferences; and (2) to showcase how a relatively simple and highly replicable approach for testing hidden selection bias can be easily adopted in educational research with complicated data and statistical model structures.

Background

Experimental Designs versus Quasi-Experiments

First proposed by Fisher, experimental designs have been viewed as the "gold standard" of research designs (1935). As Kirk stated in 1995, "an experimental design is a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan" (1995, p. 1). According to Kirk, the primary goal of an experimental design is to identify the causal relationship between independent (the assumed causes) and dependent variables (the outcomes). The key element of such a design is the utilization of randomization in the assignment of units to the treatments under study. However, due to practical and ethical reasons, experimental designs may not always be feasible in reality so researchers turn to quasi-experiments, studies without random assignment, instead.

A major disadvantage of quasi-experiments is that they potentially suffer from selection bias, a type of threat to internal validity. Internal validity in both experimental designs and quasi-experiments is about the credibility of causal inferences. It refers to "inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured" (Shadish, Cook, & Campbell, 2002, p. 53). In other words, are the differences observed between the experimental and the control groups on the outcome due primarily to the intervention? Are there any other extraneous variables that influence the outcome?

Methods to Reduce Overt Selection Bias

To reduce selection bias, researchers have done extensive work to develop appropriate statistical adjustments and research designs to minimize potential differences between treatment groups. Available statistical adjustments include, but are not limited to, regression adjustments and matching methods such as propensity score matching developed by Rosenbaum and Rubin (1983a). Moreover, research designs to address overt selection bias consist of randomized control trials, regression discontinuity, and interrupted time series, to name a few. These designs are believed to be more powerful tools in making causal arguments than statistical

Chen An, Department of Research and Evaluation, Orange County Public Schools, Orlando, FL; chen.an@ocps.net. Henry Braun, Department of Measurement, Evaluation, Statistics, and Assessment, Boston College, Chestnut Hill, MA; henry.braun@bc.edu. Mary E. Walsh, Department of Counseling Psychology, Boston College, Chestnut Hull, MA; mary.walsh@bc.edu.

adjustments. As Rubin (2008) stated, “For objective causal inference, design trumps analysis” (p. 1).

Sensitivity Analysis to Measure Hidden Bias

The statistical adjustments and research designs noted, if applied appropriately, can remove overt bias in observational studies. However, hidden bias, as embodied by unobserved characteristics that are unintentionally omitted from analytic models, may remain. Sensitivity analysis was developed to address this issue, by estimating “how much hidden bias would need to be present if hidden bias was to explain the differing outcomes in the treated and control groups” (Rosenbaum, 1991a, p. 901).

There are different approaches to assess sensitivity to hidden bias depending on the types of unobserved covariate and the outcome and the statistical tests being used, as well as the number of treatment groups. For instance, Rosenbaum and Rubin (1983b) proposed a simple technique to estimate the average effect of a treatment (one treatment group and one comparison group) on a binary outcome after adjusting for observed categorical covariates and an unobserved binary covariate U. Using a maximum likelihood estimation procedure, the difference in probabilities of the expected outcome between the treatment and the control groups (the treatment effect) can be repeatedly estimated by altering assumptions about U. These assumptions include different values of the increase in the log odds of receiving the treatment associated with $U = 1$ rather than with $U = 0$; different values of the increase in the log odds of the expected outcome under one treatment associated with $U = 1$ rather than with $U = 0$; and different proportions of participants with $U = 0$. By examining the resulting different estimated treatment effects, one can infer how extreme the assumptions about the parameters governing U must be to meaningfully change the conclusions about the treatment effect.

More sophisticated techniques have been developed by Rosenbaum (1988, 1989) to address multiple control and treatment groups. Efforts also have been made to apply sensitivity analysis to permutation tests (Rosenbaum, 1987; Rosenbaum & Krieger, 1990) and in the context of multiple regression and matched case-control studies (Rosenbaum, 1986, 1991b).

Influenced by Rosenbaum’s early work, with modifications suggested by Montgomery, Richards, and Braun (1986), Diaconu (2012) adopted a sensitivity analysis method assuming the existence of a binary unobserved variable that was related to both the binary treatment group assignment and a continuous outcome. The current study presents a sensitivity analysis following Diaconu’s approach, adapted with a larger number of simulations and a more complex statistical model to estimate the treatment effect over multiple years. Diaconu’s approach was followed primarily because her work examined the effectiveness of a quasi-experimental intervention in a multilevel modeling context, similar to the current study.

Methods

Treatment

It has long been recognized that life outside of school has consequences for achievement in school, especially for students growing up poor (Coleman et al., 1966; Dearing, 2008; Harrington, 1962). City Connects is a comprehensive student support intervention designed in response to the recognition that social and behavioral factors in the context of academic learning may impede students’ ability to benefit from instruction (Walsh & Brabeck, 2006). Its mission is to help children succeed and thrive by connecting each student in a school with a tailored set of prevention, intervention, and enrichment services. Initially serving kindergarten to fifth-grade students in an urban district in Massachusetts, City Connects has been extended to prekindergarten through high school and to other districts across the country.

A full-time masters’ trained licensed school counselor or social worker in each school, the coordinator, is at the core of the intervention. In the fall of each school year, each classroom teacher and the coordinator discuss strengths and needs of *every* student in the class relative to four developmental domains—academic, social/emotional/behavioral, family, and health. For children identified as having intensive needs, coordinators arrange additional reviews with a wider team of professionals to discuss specific goals and strategies for that student.

Next, coordinators develop tailored plans for each student, identifying supports that could promote student strengths and meet student needs in the four domains. Using a computer-based tool designed for the intervention, they find specific school- or community-based providers based on factors such as service type(s), geographical location, and transportation requirements. Examples of services include before- and after-school programs, sports programs, mental health counseling, mentoring, tutoring, attendance support, health screenings, family counseling, arts and music programs, and food or clothing donations. Coordinators connect children and their families with service providers, monitor service quality and fit, maintain partnerships with community providers, and help to coordinate activities of agencies working within the school. A documented, standardized set of practices, oversight mechanisms, and fidelity tools guide implementation across sites and services.

Past quasi-experimental research has demonstrated beneficial effects of City Connects (Dearing, Sibley, Lee-St. John, Raczek, & Walsh 2016; Walsh et al., 2014). The evaluation of the intervention has been designed as a quasi-experiment, because participating schools were identified by the district to receive the full program treatment. Also, because serving all students within a school is a critical feature of the intervention, individual students could not be assigned to treatment or control conditions within schools. Because City Connects is assigned at the school level and the intervention carried out in an individualized/tailored way for each student, analyses rely on multilevel modeling techniques with intervention assignment at the school level and variation in number of years of treatment, or “dosage,” at the student level.

Measures

Demographics. Student demographic characteristic variables included gender, ethnicity (categorized as African American, Asian, Hispanic, White, or other non-White minority), poverty status (measured by eligibility for free- or reduced-price lunch), English language proficiency (district classification), special education status, immigrant student status, academic mobility (number of school changes for the student), distance (in miles) from home to school zip code, and student age in the current year.

Teacher-rated baseline measures. In this school district, elementary school report card scores were assigned in a standard way. First-grade report card scores used in propensity score model development were calculated as the sum of values (1–4) assigned to multiple reading, mathematics, writing, behavior, academic work habits, and classroom effort items to produce a single total score for each subject.

Academic outcomes. The Massachusetts Comprehensive Assessment System (MCAS) was a statewide standardized battery required of all public school students in Massachusetts during the years of this study. The outcomes of interest were MCAS mathematics and English language arts (ELA) scores. All MCAS raw scores were converted into z scores by subject, by grade, and by school year using the means and standard deviations of the comparison group.

Sample

The analytic sample included approximately 9,000 students in a large, urban, Massachusetts public school district who entered kindergarten in school years 2000–2001 through 2006–2007 and reached at least Grade 6 by the 2012–2013 academic year. The treatment sample comprised all students who attended one of 13 schools that implemented City Connects ($N = 1{,}500$). The comparison sample included all students who attended one of 77 non–City Connects schools in the district ($N = 7{,}500$). The sample excluded students with severe special needs who required instruction in substantially separate classrooms, although other special education students were included. Sample members were required to have baseline achievement measures (about half of the students in the sample had valid baseline scores[1]).

To examine trends over time, student records were followed over the 3 years of middle school. All students that reached Grade 6 in the district by the 2012–2013 academic year were included in the Grade 6 models; among these students, those who continued enrolling in within-district middle schools and never switched schools were included in the Grade 7 models; and those who stayed in the district and never switched schools for the entire 3 years of middle school were included in the Grade 8 models. The sample size dropped to 5,000 in Grade 7 and subsequently to 3,500 in Grade 8. This degree of attrition was primarily due to students not yet reaching the outcome grade and partly due to students transferring to other schools. The former is a design issue and may not be a serious threat. The latter, however, demands some further study.

Propensity Score Weighting

Propensity score weights were generated first to remove overt selection bias. Results showed that applying such weights did significantly reduce overt selection biases. In other words, the treatment and the comparison groups were nearly statistically equivalent in terms of baseline (Grade 1) characteristics after applying the propensity score weights. The treatment group started with a relatively more disadvantaged population of students than the comparison group: a larger proportion of students were bilingual or born outside of the United States; they usually lived farther away from school; and more of them struggled with poverty and suffered low achievement. The treatment group also included more Asian students (24.3% versus 6.2%). These imbalances were removed by the propensity score weighting procedure, gauged by both standardized bias and the p-value methods, the standard statistics to check covariate balance between the treatment and the comparison groups.

Baseline Estimation of Treatment Effects

In order to estimate the overall treatment effect of the elementary school intervention on student academic achievement in middle school, a two-level hierarchical linear model that took into account elementary school clusters (the most recent intervention or comparison elementary schools that students attended), together with statistical adjustments of students' baseline achievement, demographic characteristics,[2] and differential middle school effectiveness,[3] was built in each middle school grade and for each MCAS subject. A binary treatment effect indicator was placed at the school level for estimation. The full model is expressed by Equation 1 and $\hat{\gamma}_{01}$ was the estimated treatment effect.

Level 1 (student level):

$$Y^{j}_{ik} - Z^{j}_{ik} = \beta_{0k} + \beta_{1k}X^{j}_{1ik} \ldots + \beta_{qk}X^{j}_{qik} + \sum_{m=1}^{5} \theta_{mk} Dosage^{j}_{mik} + r^{j}_{ik}.$$

Level 2 (school level):

$$\begin{aligned} \beta_{0k} &= \gamma_{00} + \gamma_{01}\, EDose_{k} + u_{0k} \\ \beta_{1k} &= \gamma_{10} + u_{1k} \\ &\ldots \\ \beta_{qk} &= \gamma_{q0} + u_{qk} \\ \sum_{m=1}^{5} \theta_{mk} &= \sum_{m=1}^{5} (\varphi_{m0} + \psi_{mk}), \end{aligned} \tag{1}$$

with denotations as follows:

- i denotes students within elementary schools, k denotes the last elementary school attended, and j denotes middle schools;
- Y^{j}_{ik} is the academic outcome measure (MCAS Mathematics or ELA) in one of the middle school grades for student i in the last elementary school k who then went to middle school j;
- Z^{j}_{ik} is achievement adjustment score for middle school j attended by student i in elementary school k;
- X^{j}_{1ik} to X^{j}_{qik} are q student-level covariates for student i in the last elementary school k who then went to middle school j;
- $\sum_{m=1}^{5} Dosage^{j}_{mik}$ represents a series of dummy variables indicating the number of years spent in the treatment elementary schools, where $m = 1, 2, \ldots\ 5^4$;
- β_{0k} is the mean of the outcome measure for the last elementary school k, adjusted for middle school achievement (Z^{j}_{ik}), q covariates (X^{j}_{1ik} to X^{j}_{qik}), and $\sum_{m=1}^{5} Dosage^{j}_{mik}$;
- β_{1k} to β_{qk} are the regression coefficients for the last elementary school k, associated with q student-level covariates (X^{j}_{1ik} to X^{j}_{qik});

- $\sum_{m=1}^{5} \theta_{mk}$ are regression coefficients for the last elementary school k, associated with $\sum_{m=1}^{5} Dosage_{mik}^{j}$;
- r_{ik}^{j} is the random error (or residual) at level 1, where $r_{ik}^{j} \sim N(0, \sigma^2)$ and σ^2 is the variance of the student-level residuals;
- $EDose_k$ is a dummy variable indicting treatment membership in elementary school, with 1 for treatment schools, and 0 for comparison schools;
- γ_{01} is the estimated treatment effect;
- γ_{00} is the intercept at level 2, which is the adjusted mean achievement for comparison elementary schools (i.e., when $EDose_k = 0$);
- γ_{10} to γ_{q0} are constants indicating the means of the q regression coefficients across all last elementary schools;
- $\sum_{m=1}^{5} \varphi_{m0}$ are constants indicating the mean values of $\sum_{m=1}^{5} \theta_{mk}$ across all last elementary schools;
- u_{0k} to u_{qk} are random effects at level 2, where $u_{vk} \sim N(0, \tau_v)$ $(v = 0, 1, \ldots, q)$ and τ_v is the variance of the school-level residuals for $\{\beta_{vk}\}_k$ $(v = 0, 1, \ldots, q)$; and
- $\sum_{m=1}^{5} \psi_{mk}$ are random effects at level 2, where $\psi_{mk} \sim N(0, \tau_m')$ $(m = 1, 2, \ldots, 5)$.

Sensitivity Analysis

Hypothesizing the unobserved variable U. To start with, one needs to hypothesize a real but unobserved variable U that bears some relationship with both treatment assignment and the outcome. For the sake of easier interpretation, one would like to require both relationships to be positive because the estimated positive treatment effect will be inflated if there is a confounding of treatment and selection. By including an unobserved variable that is positively related to the treatment assignment (selection), estimates of the treatment effect will be smaller. Meanwhile, if this unobserved variable is also positively related to the outcome, it will further reduce the estimated treatment effect.

In this study, *parental involvement*, a dummy variable indicating whether parents are involved with their children's education, was assumed to be the unmeasured variable *U*. *Parental involvement* was chosen because literature suggested that it has positive relationships with both school choice and academic achievement (Epstein, 1995; Fan & Chen, 2001). It is necessary to mention that any existing yet unobserved variable that bears the assumed relationships could have been used. This variable can be directly collected from a parent questionnaire asking whether parents believe they are involved with their children's education or from a teacher survey asking teachers about their impression of each student's parental involvement. However, a measure of *parental involvement* was not available in this study.

Assumptions about U. As described by Rosenbaum (1988), Montgomery, Richards, and Braun (1986), and Diaconu (2012), two assumptions need to be made to define the conditions of the simulated U. The first assumption (Assumption 1) deals with how U is related to Z, the binary treatment status. The conditional probability of U taking any value of u given Z taking any value of z is expressed mathematically as $\Pr(U = u \mid Z = z)$, where $u = 1$ indicates high *parental involvement*, $u = 0$ indicates low *parental involvement*, $z = 1$ for the treatment group, and $z = 0$ for the comparison group. There are four sets of conditional probabilities of U given Z: the conditional probability of $U = 0$ given $Z = 0$ is denoted as $\pi_{0|0}$; the conditional probability of $U = 1$ given $Z = 0$ is denoted as $\pi_{1|0}$; the conditional probability of $U = 0$ given $Z = 1$ is denoted as $\pi_{0|1}$; and the conditional probability of $U = 1$ given $Z = 1$ is denoted as $\pi_{1|1}$.

All these four parameters ($\pi_{1|1}$, $\pi_{1|0}$, $\pi_{0|0}$, and $\pi_{0|1}$) can take on any values between 0 and 1 (because they are probabilities so the range is 0–1) as long as they meet the requirement of Assumption 1. Large $\pi_{1|1}$–$\pi_{1|0}$ or $\pi_{0|0}$–$\pi_{0|1}$ corresponds to the existence of a strong selection bias since U is strongly associated with Z. On the contrary, small $\pi_{1|1}$–$\pi_{1|0}$ or $\pi_{0|0}$–$\pi_{0|1}$ indicates that the problem of selection bias may not be so severe.

Note that the conditional probability of U taking one value of u is dependent on the conditional probability of U taking the other value of u given Z taking the same value of z.

Therefore, one can just focus on the relationship represented by one of the two inequalities ($\pi_{1|1} > \pi_{1|0}$ or $\pi_{0|0} > \pi_{0|1}$). Assumption 1 is simplified as $\pi_{1|1} > \pi_{1|0}$. The range of $\pi_{1|0}$ and $\pi_{1|1}$ was set to be .2–.8 with .15 as a basic incremental unit. To satisfy $\pi_{1|0} < \pi_{1|1}$, 10 pairs of possible values of $\pi_{1|0}$ and $\pi_{1|1}$ were generated.

The second assumption deals with how U is related to the outcome of interest. A reasonable hypothesized relationship between parental involvement and academic achievement is that they are positively related. In other words, the regression coefficient associated with U should be positive. Following Rosenbaum's approach (1986), values were set based on empirical results obtained from baseline estimation models. Possible values of the regression coefficient associated with U were (1) the largest regression coefficients associated with student-level demographic covariates in the outcome models; and (2) the ones associated with prior or baseline achievement adjustments.

The predetermined regression coefficient associated with each set of the simulated U was set equal to .3 based on the empirical results of Equation 1. Among all the positive (estimated) regression coefficients indicating the relationships between the corresponding variables and the outcomes, .3 is considered relatively strong and influential. Using the outcome model predicting Grade 6 MCAS ELA as an example, the relationship between student immigrant status (variable *foreign born*) and sixth grade ELA scores is .26, which is the largest positive regression coefficient[5] in that model. The largest regression coefficient associated with prior achievement is .12 between Grade 1 reading report card scores and Grade 6 MCAS ELA. Thus, if the regression coefficient of the hypothesized U has a value of .30, this U would be more influential than student immigrant status and any of the prior achievement measures in predicting the outcome. Whether or not such a strongly influential U exists in reality is questionable. However, this conservative approach of assuming a relatively strong relationship will result in reductions of the estimated treatment effects. If the reestimated effects are still significant, then this will strengthen the argument that the intervention indeed produced consistent and positive impacts on the outcomes of interest, despite the presence of selection bias.

General procedure of sensitivity analysis. Random samples of U can then be drawn through Monte Carlo simulation, which utilizes random numbers in the simulation algorithm (Kennedy, 2003). The principle behind this is that "the behavior of a statistic in random samples can be assessed by the empirical process of actually drawing lots of random samples and observing this behavior" (Mooney, 1997, p. 2). As described by Mooney (1997), Monte Carlo simulation is used to create a pseudo-population which possesses key mathematical properties that make it resemble samples of data drawn from the true population. Then multiple trials are drawn from the pseudo-population to conduct statistical analysis to investigate how the procedure behaves across trials.

We reestimated the treatment effects with U included as a predictor. Equation 2 is the same as Equation 1 except that it has one set of the newly simulated U_{ik}^{j} and a predetermined regression coefficient β_{Uk}^{j} adjusted to the outcome. The estimated treatment effect with U included is $\hat{\gamma}'_{01}$. The difference between $\hat{\gamma}'_{01}$(obtained from Equation 2) and $\hat{\gamma}_{01}$(obtained from Equation 1) is the hidden bias to the treatment estimate if U is not included in the model.

Level 1 (student level):

$$Y_{ik}^{j} - Z_{ik}^{j} - \beta_{Uk}^{j} U_{ik}^{j} = \beta'_{0k} + \beta'_{1k} X_{1ik}^{j} \ldots + \beta'_{qk} X_{qik}^{j} + \sum_{m=1}^{5} \theta'_{mk} Dosage_{mik}^{j} + r_{ik}^{\prime j}.$$

Level 2 (school level):

$$\begin{aligned} \beta'_{0k} &= \gamma'_{00} + \gamma'_{01} EDose_k + u'_{0k} \\ \beta'_{1k} &= \gamma'_{10} + u'_{1k} \\ &\ldots \\ \beta'_{qk} &= \gamma'_{q0} + u'_{qk} \end{aligned} \tag{2}$$

$$\sum_{m=1}^{5} \theta'_{mk} = \sum_{m=1}^{5} \left(\varphi'_{m0} + \psi'_{mk}\right),$$

where U_{ik}^{j} is one set of the simulated values for the unobserved variable U for student i in elementary school k with adjustment on middle school$j$; and β_{Uk}^{j} is the predetermined regression coefficient associated with U_{ik}^{j} .

For each of the 10 pairs of conditional probabilities $\pi_{1|0}$ and $\pi_{1|1}$, we simulated U and reran the models 100 times and an average estimated treatment effect across the 100 trials was calculated. Then the average treatment effects and the corresponding one-sided 90% confidence intervals[6] were shown to examine the extent of such bias.

Results

Original Estimated Treatment Effects

We report the estimated treatment effects and effect sizes in Table 1. According to Table 1, original effect sizes in Grades 6–8 are quite large in magnitude with values of .62, .42, and .67 for mathematics and .41, .29, and .44 for ELA, indicating that the estimated treatment effects of the intervention on middle school academic outcomes are not only statistically significant but also practically significant: the treatment group outperformed the comparison group by at least a quarter of a standard deviation (i.e., for the outcome of MCAS ELA in Grade 7) and at most nearly two-thirds of a standard deviation (i.e., for the outcome of MCAS mathematics in Grade 8).

Table 1. Estimated Treatment Effects ($\hat{\gamma}_{01}$) in Middle School

	Coef.	SE	*p*-Value	Effect Size
MCAS Mathematics				
Grade 6	.67	.25	.009	.62
Grade 7	.38	.16	.022	.42
Grade 8	.63	.16	.000	.67
MCAS ELA				
Grade 6	.43	.17	.016	.41
Grade 7	.25	.08	.004	.29
Grade 8	.38	.09	.000	.44

Note: Following the approach suggested by the What Works Clearinghouse (2011, F. 9), effect sizes were computed as Hedges' *g*, the ratio of the estimated treatment effect to the unadjusted pooled within-group *SD* using the formula below:

$$g = \frac{\omega\gamma}{\sqrt{\frac{(n_i-1)s_i^2+(n_c-1)s_c^2}{n_i+n_c-2}}}, \tag{5}$$

where γ is the HLM coefficient for the intervention's effect; ω is the small-sample bias corrector calculated as $1 - 3/(4df - 1)$ (Hedges, 1981, p. 114), with *df* being the number of degree of freedom used to estimate the standard deviation; s_i and s_c are the unadjusted standard deviations of the treatment and the comparison groups, separately; and n_i and n_c are the sample sizes of the two groups.

To contextualize such effect sizes, Hill, Bloom, Black, and Lipsey (2008) developed an empirical benchmark regarding effects of similar types of educational interventions using meta-analysis. They examined 36 randomized assignment studies of educational interventions with middle school achievement data as outcome measures and identified an unweighted average effect size of .51 with a standard deviation of .33. Based on this benchmark, we believe our intervention is quite successful on mathematics and moderately effective on ELA.

Results Predicting MCAS Mathematics

Table 2 presents the estimated treatment effects of the elementary school intervention on MCAS mathematics in Grades 6–8 ($\hat{\gamma}'_{01}$) when the outcome models included different sets of U. In addition, the ones estimated from models that excluded U ($\hat{\gamma}_{01}$), together with the corresponding standard errors and the one-sided 90% confidence intervals, are shown at the bottom of the table for comparison.

As shown in Table 2, the original estimated treatment effects are .65, .38, and .56 in Grades 6–8, respectively. By including the unobservable U, the estimated treatment effects shrunk slightly: the ranges of these estimates are from .46 to .60, from .20 to .34, and from .38 to .52 in Grades 6–8, respectively. However, since all these estimates fall within the one-sided 90% confidence intervals of the original ones, one can conclude that the estimated treatment effects are reasonably robust to the presence of the type of hidden bias specified in this study.

The range of $\pi_{1|0}$ and $\pi_{1|1}$ was set to be .20–.80 with .15 as a basic incremental unit. Initially, it was assumed that the

Table 2. The Estimated Treatment Effects ($\hat{\gamma}'_{01}o$) of RQ1 With Different Sets of *U* Included in the Outcome Models Predicting MCAS Mathematics

			$\hat{\gamma}'_{01}$		
U	$\pi_{1\|0}$	$\pi_{1\|1}$	Grade 6	Grade 7	Grade 8
u_1	.20	.35	.60	.34	.51
u_2	.20	.50	.55	.29	.47
u_3	.20	.65	.51	.25	.43
u_4	.20	.80	.46	.20	.38
u_5	.35	.50	.60	.34	.51
u_6	.35	.65	.56	.29	.47
u_7	.35	.80	.51	.25	.43
u_8	.50	.65	.60	.34	.52
u_9	.50	.80	.55	.29	.47
u_{10}	.65	.80	.60	.34	.52
		$\hat{\gamma}_{01}$	.65	.38	.56
		Effect Size	.60	.42	.59
		SE	.15	.14	.15
		One-Sided 90% CI	[.39, .65]	[.16, .38]	[.31, .56]

Note: Notice that the original estimated treatment effects are slightly lower than those reported in Table 1. This is due to (1) the change of software from HLM to Stata in order to run multiple trials; and (2) propensity score weights were removed (comparisons were done and no significant differences were observed) because Stata does not allow such weights with restricted maximum likelihood estimation.

magnitude of these conditional probabilities represented the severity of selection bias. For instance, u_1 ($\pi_{1|0} = .20$ and $\pi_{1|1} = .35$) represented relatively mild selection bias; while u_{10} ($\pi_{1|0} = .65$ and $\pi_{1|1} = .80$) represented relatively strong selection bias. However, the empirical results suggested that it was the difference between the values of each pair of conditional probabilities that determined the severity of such bias.

As shown in Table 2, the estimated treatment effects are approximately the same in each grade if the difference in value between the corresponding conditional probabilities is the same. For instance, $\hat{\gamma}'_{01}$ is approximately .60, .34, and .51 in Grades 6–8, respectively, for u_1 ($\pi_{1|0} = .20$ and $\pi_{1|1} = .35$), u_5 ($\pi_{1|0} = .35$ and $\pi_{1|1} = .50$), and u_{10} ($\pi_{1|0} = .65$ and $\pi_{1|1} = .80$) because the difference in value between the two conditional probabilities is .15 for all these three pairs. Thus, u_4 ($\pi_{1|0} = .20$ and $\pi_{1|1} = .80$) represents relatively strong selection bias, and u_1 ($\pi_{1|0} = .20$ and $\pi_{1|1} = .35$), u_5 ($\pi_{1|0} = .35$ and $\pi_{1|1} = .50$), and u_{10} ($\pi_{1|0} = .65$ and $\pi_{1|1} = .80$) represent relatively mild selection bias of the same degree.

To further establish this point, Table 3 presents detailed results of the outcome models predicting MCAS mathematics in Grade 6 when the unobservable *U* was included. *U* was simulated based on two pairs of conditional probabilities of *U* given *Z*: u_1 ($\pi_{1|0} = .20$ and $\pi_{1|1} = .35$) and u_{10} ($\pi_{1|0} = .65$ and $\pi_{1|1} = .80$). As can be seen, the results obtained using u_1 and u_{10} are nearly identical.

Table 3. Results Comparison Using Different Pairs of Conditional Probabilities of *U* Given *Z* (Outcome Models Predicting MCAS Mathematics in Grade 6 with the Unobserved *U* Included)

	u_1			u_{10}		
	$\pi_{1\|0} = .20$ and $\pi_{1\|1} = .35$			$\pi_{1\|0} = .65$ and $\pi_{1\|1} = .80$		
Fixed Effects	Coef.	SE	*p*-Value	Coef.	SE	*p*-Value
Intercept	1.30	.20	.000	1.11	.20	.000
EDose (Ever Received the Treatment)	.60	.15	.000	.61	.15	.000
1-Year vs. 6-Year Dosage	−.55	.08	.000	−.58	.08	.000
2-Year vs. 6-Year Dosage	−.35	.08	.000	−.36	.08	.000
3-Year vs. 6-Year Dosage	−.29	.09	.001	−.31	.09	.000
4-Year vs. 6-Year Dosage	−.36	.08	.000	−.37	.08	.000
5-Year vs. 6-Year Dosage	−.13	.09	.167	−.16	.09	.077
Male	.03	.02	.184	.02	.02	.377
is_Black	−.19	.04	.000	−.19	.04	.000
is_Asian	.37	.05	.000	.37	.05	.000
is_Hispanic	−.07	.04	.081	−.07	.04	.101
is_Other	−.12	.08	.124	−.11	.08	.177
Bilingual	−.05	.03	.059	−.05	.03	.087
Special Needs 2	−.20	.04	.000	−.20	.04	.000
Special Needs 3	−.59	.04	.000	−.59	.04	.000
Reduced Lunch	−.16	.07	.020	−.17	.07	.017
Free Lunch	−.26	.04	.000	−.25	.04	.000
Foreign Born	.15	.04	.000	.15	.04	.000
RC_Reading_gr1	.06	.02	.002	.06	.02	.002
RC_Math_gr1	.17	.02	.000	.17	.02	.000
RC_Writing_gr1	.04	.02	.040	.04	.02	.036
RC_WorkHabit_gr1	.09	.02	.000	.09	.02	.000
RC_Behavior_gr1	−.01	.02	.588	−.01	.02	.429
RC_Effort_gr1	.00	.02	.957	.00	.02	.906
Age_gr1	−.16	.03	.000	−.15	.03	.000
Distance from School_gr1	.00	.01	.490	.00	.01	.544
# School Moves_gr1	.00	.02	.849	.02	.02	.474

Table 4. The Estimated Treatment Effects ($\hat{\gamma}'_{01}$) of RQ1 With Different Sets of *U* Included in the Outcome Models Predicting MCAS ELA

			$\hat{\gamma}'_{01}$		
U	$\pi_{1\|0}$	$\pi_{1\|1}$	**Grade 6**	**Grade 7**	**Grade 8**
u_1	.20	.35	.35	.16	.31
u_2	.20	.50	.31	.11	.27
u_3	.20	.65	.26	.07	.23
u_4	.20	.80	.22	.02	.18
. . .					
		$\hat{\gamma}_{01}$	.40	.20	.36
		Effect Size	.38	.23	.42
		SE	.14	.11	.13
		One-Sided 90% CI	[.17, .40]	[.01, .20]	[.15, .36]

Note: The resulting estimated treatment effects with u_5–u_{10} included displayed a repetitive pattern due to the way the sets of *U* were chosen (i.e., using .15 as an incremental unit); therefore, they were omitted from this table.

Results Predicting MCAS ELA

Table 4 presents the estimated treatment effects of the elementary school intervention on MCAS ELA in Grades 6–8 ($\hat{\gamma}'_{01}$) when the outcome models included different sets of *U*. In addition, the ones estimated from models that excluded *U* ($\hat{\gamma}_{01}$), together with the corresponding standard errors and the 90% confidence intervals, are shown at the bottom of the table for comparison.

As shown in the Table 4, the original estimated treatment effects are .40, .20, and .36 in Grades 6–8, respectively. By including the unobservable *U*, the estimated treatment effects shrunk slightly: the ranges of these estimates are from .22 to .35, from .02 to .16, and from .18 to .31 in Grades 6–8, respectively. The lowest estimated treatment effect with *U* included is .02 for the outcome model predicting Grade 7 ELA with $\pi_{1|0} = .20$ and $\pi_{1|1} = .80$ (the presence of a relatively strong selection bias), which is quite small. However, because all these ranges fall within the one-sided 90% confidence intervals of the original estimates, one can conclude that the estimated treatment effects are robust to the presence of hidden bias specified in this study.

Discussion

Overview

To recap, this study conducted a sensitivity analysis to assess the robustness of the estimated treatment effects of an elementary school intervention on middle school academic outcomes to the presence of hidden selection bias. Sets of binary variable *U* that met two key assumptions were randomly generated using Monte Carlo simulation. The first assumption deals with how *U* is related to *Z*, the indicator of treatment assignment: by altering the magnitude of the conditional probabilities of *U* given *Z*, 10 sets of such probabilities were chosen. The second assumption predetermines the magnitude of the relationship between *U* and the outcomes. Based on empirical results, .3 was chosen as the magnitude of the corresponding regression coefficient, because it was the largest regression coefficient associated with (observed) covariates across subjects and grade levels. The newly generated sets of *U* were then added into the outcome models to examine how the estimated treatment effects would be affected by the inclusion of *U*. This procedure was repeated 100 times and the resulting treatment effect estimates were averaged over the 100 trials.

The results showed that the estimated treatment effects for both MCAS mathematics and ELA were reduced slightly with the inclusion of *U*; however, the fact that they still fell within the one-sided 90% confidence intervals of the original ones indicated only a mild sensitivity to hidden bias. In addition, the higher the strength of the selection bias, as partly indicated by the mathematical difference between each pair of the conditional probabilities of *U* given *Z*, the smaller the estimated treatment effects.

Some may argue that the large estimated effects were a result of attrition bias. Since the treatment students started quite low at the baseline, the students struggling the most might have already dropped out of the treatment group by the time of the outcome assessments. Basic descriptive statistics suggested that it is not the case: 15.6% of students in the treatment group dropped out as compared to 16.0% in the comparison group, which is not statistically different.

In addition, we did notice that there is a consistent drop in estimated treatment effects in Grade 7 for both the mathematics and the ELA models. This may be related to the opportunity for students in this district to apply to selective public high schools in Grade 7. Acceptance is based entirely on students' academic performance and test scores. Research on estimating the impact of City Connects on selective high school admissions is still under way.

Limitations

Regarding sensitivity analysis. More studies are needed to improve upon the sensitivity analysis conducted in this study. First, in addition to simply providing descriptive statistics as in Tables 2 and 4, inferential statistics to measure the severity of hidden bias should be developed. Second, the current study only simulated 100 trials for each of the predetermined 10 sets of unobserved *U* due to the lack of computational resources. It can be argued that the number of trials is far from sufficient. Byrne (2013) suggested various minimum number of model runs to achieve desired confidence intervals for model fitting to proportion, recommending that approximately 250 repetitions for each set seem to justify the stability of the corresponding statistical model. Therefore, with higher computational resources available, more trials should be simulated to get more accurate estimates.

Third, the resulting estimated treatment effects with *U* included displayed a repetitive pattern due to the way the sets of *U* were chosen (i.e., using .15 as an incremental unit). In the future, random pairs of *U* should be generated if they meet the corresponding assumptions about *U* given *Z*. Furthermore, a stratified sampling method can be used when drawing these random pairs: the random selection will occur within each stratum that represents mild, medium, and high sensitivity (i.e., the difference in values of the conditional probabilities of *U* given *Z* is small, medium, and large). Moreover, multiple predetermined regression coefficients associated with *U* should be chosen to examine the impact of altering the magnitude of the relationship between *U* and the outcome on the estimated treatment effects. Finally, simulated *U* in this study was addressed at the student level only. There is no known research providing guidelines for sensitivity analysis in the context of multilevel models and further research is needed.

Regarding causal argument. Many other practical and methodological assumptions need to be established when using multilevel models to support causal arguments. To start with, every covariate needs to be collected without error. This study obtained student-level covariates directly from the local school district and used multiple years of data to confirm time-invariant variable values, which provides some confidence in their accuracy. However, for time-varying variables we relied on the districts' reporting system. Some variables were not available in the early years and others might not be collected in a consistent manner over years. Program-level characteristics were not included in the current analysis. Without including such factors (e.g., the number of services, the level of monitoring services, to name a few), potential hidden bias may present. However, it is unlikely that these factors would influence both the selection and the outcome achievement more than student characteristics and prior achievement measures did.

School-level characteristics[7] were derived from state published data files. Including them into the outcome models resulted in no change to the estimated treatment effects on ELA but a slight increase to those in mathematics. For the most conservative estimate of the treatment effects to compare simulated data against, we decided to not include them. In addition, prior achievement measures should meet professional standards of reliability and validity as the outcome measures do (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014). Correlations between report card scores and MCAS scores in the same domains—that is, reading and writing report cards with MCAS ELA and math report cards with MCAS math—were moderately high in this sample (.64–.70); a strong but not absolute relationship, pointing to convergent validity. Correlations for scores across domains were lower than same-subject correlations in every instance at each grade, evidence of discriminant validity. Nonetheless, measurement characteristics of the teacher ratings have not been evaluated to the same extent as MCAS scores.

Moreover, statistically speaking, Reardon and Raudenbush summarized six assumptions needed to draw unbiased causal inferences: (1) manipulability; (2) no interference between units; (3) interval scale metric; (4) homogeneity of effect; (5) strong ignorable treatment assignment; and (6) functional form (2009, p. 18). Only one of them (i.e., strong ignorability assumption) was tested through sensitivity analysis.

First, manipulability means that each student can attend any school and each student has a nonzero probability of attending any school. In reality, this assumption is often violated. For instance, in this studied district, student enrollment assignment is dependent on (1) if schools are located in the zone in which students live; (2) if schools in other zones are within their walk zone; or (3) if schools are citywide K-8 and middle schools that are open to all students. Additionally, some schools require an interview or assessment to attend. Therefore, students cannot attend any schools they want.

Second, under this context, no interference between units means that there is no peer effect—the assignment of one student to one school is not dependent on the school assignment of other students. If student composition affects instructional practices and curricula and thus affects student learning, or if student composition affects school recruiting students and teachers, then this assumption is violated.

Third, interval-scale metric assumption requires the observed outcome to be on an interval scale. Unfortunately, statewide standardized tests such as MCAS do not meet this assumption. Unlike physical quantities such as weights and heights, neither Mathematics nor ELA is a unidimensional construct that can be placed on a continuum with equal intervals and this dimensional mix changes from grade to grade. Item response theory (IRT) models need to be constructed to obtain genuine interval scales.

Fourth, homogeneity means that school effect is constant across students who attend the given school. If a teacher or a school caters instructions to a special population of students, then this teacher or school will be more effective to such students than other teachers or schools without such a student composition. Last, common support/functional form assumption requires modeling the function form correctly for students who are not present in a given school.

In this study, as explained above, the assumptions of manipulability and no interference between units are implausible in the current school system. Item-level responses are required to examine the interval-scale metric assumption. Unfortunately, the studied district did not provide such information. Homogeneity and function form assumptions were discussed in details in Reardon and Raudenbush's (2009) study. Therefore, due to its strong relevance to causal argument, as well as the availability of corresponding statistical testing techniques, the consequences of violations to the strongly ignorable treatment assignment assumption were assessed, assuming all the other five assumptions were held. Future studies are needed to examine all the applicable assumptions to justify the causal argument that the intervention leads students to prosper academically in the long run.

Final Remarks

Making causal inferences for a quasi-experiment is extremely difficult. Sensitivity analysis approaches, which address hidden selection bias, thus have gained popularity. This study serves as an introduction to sensitivity analysis in a simple but practical form, applied to a complex multiyear and multicluster project studying the effectiveness of an intervention. Computer capacities have prevented the study from simulating a larger number of trials based on industry standards; however, with the availability of the Linux Cluster or similar computing systems to handle heavy computations, researchers will be able to conduct a full-scale sensitivity analysis that generates more convincing evidence to support a causal claim. Furthermore, the study is only the first step into a very complicated world of testing hidden bias. Collective wisdom in strengthening the approach to causal argument, either by examining existing sensitivity analysis or by developing new approaches, is greatly needed.

Notes

[1] Percent missing prior achievement was 55.4% for the comparison group and 44.4% for the treatment group for the Grade 6 model. Similar numbers were observed for models in later grades.

[2] Baseline achievement includes Grade 1 fall report card scores in reading, mathematics, writing, behavior, work habits, and effort; while baseline demographic characteristics include gender, race, English language learner status, special education status, reduced or free lunch status, student immigrant status, age when starting Grade 1, the number of school moves when starting Grade 1, and home distance to school in miles when starting Grade 1.

[3] To accurately estimate the treatment effects, it is important to take into account the cross-classified nested structure of the data: students

attended different elementary schools, some of which were implementing the intervention; these students then progressed to different middle schools. To adjust for differential middle school effectiveness, a two-phase analysis was conducted for each subject and in each grade. In the first phase, using comparison students only, a two-level hierarchical linear model, which took into account middle school clusters and adjusted for students' MCAS ELA and mathematics scores in Grade 5 and the same set of student characteristics, was built to predict the outcomes of interest. A reasonable concern is that if there is a positive treatment effect on academic achievement, consequently graduates of the treatment elementary schools will be more likely to attend better middle schools. As a result, adjusting for middle school effectiveness using both the treatment and the comparison students will remove this positive effect undiscriminatingly, if it exists, and thus underestimate the treatment effect in the outcome models. In other words, it is believed that treatment elementary schools should take credit for sending their graduates to better middle schools and this effect should be reserved when adjusting middle school effectiveness in an attempt to differentiate the contribution of elementary schools on achievement from that of middle schools. To address this concern, only comparison students were used to estimate middle school effectiveness. The estimated adjusted mean of each middle school was then subtracted from the outcome scores of the analytic models described in the main text as the adjustment for differential middle school effectiveness.

[4] Since the intervention serves kindergarten to fifth grades in elementary school, the maximum years of receiving the intervention is 6 years. Using students who received 6 years of the intervention as the reference group, the number of dosage dummy variables in the equation will be $6 - 1 = 5$.

[5] Sampled immigrant students were enrolled in the public school system at early grades (having their Grade 1 report card scores indicating that they immigrated quite early). Six or more years of formal schooling and targeted language intervention programs is more likely to cancel out the generally conceived negative influence of foreign-born status, especially for some minority groups.

[6] One-sided confidence intervals were used instead of the more widely used two-sided ones. This is because by adding an influential variable U that is positively related to the outcome, the estimated treatment effect will decrease. In this case, it makes more sense to use one-sided confidence intervals.

[7] School-level characteristics that were tested included % English language learners, % low income, % students with disabilities, student/teacher ratio, school size, average class size, and students per computer.

References

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Byrne, M. D. (2013, July). *How many times should a stochastic model be run? An approach based on confidence intervals*. Paper presented at the 12th International Conferences on Cognitive Modelling, Ottawa, Canada.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: National Center for Educational Statistics.

Dearing, E. (2008). The psychological costs of growing up poor. *Annals of the New York Academy of Sciences*, *1136*, 324–332.

Dearing, E., Sibley, E., Lee-St. John, T., Raczek, A., & Walsh, M. (2016). Can community and school-based supports improve the achievement of first-generation immigrant children attending high-poverty schools? *Child Development*, *87*, 883–897.

Diaconu, D. V. (2012). *Modeling science achievement differences between single-sex and coeducational schools: Analyses from Hong Kong, SAR and New Zealand from TIMSS 1995, 1999, and 2003* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses (Accession Order No. [UMI3521765]).

Epstein, J. L. (1995). School/family/community partnerships: Caring for the children we share. *Phi Delta Kappan*, *76*, 701–712.

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review*, *13*(1), 1–22.

Fisher, R. A. (1935). *Design of experiments*. London, UK: Macmillan.

Harrington, M. (1962). *The other America: Poverty in the United States*. New York, NY: Simon & Schuster.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge, MA: MIT Press.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Montgomery, M. R., Richards, T., & Braun, H. I. (1986). Child health, breast-feeding, and survival in Malaysia: A random-effects logit approach. *Journal of the American Statistical Association*, *81*(394), 297–309.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousand Oaks, CA: Sage Publications.

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

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*(3), 207–224.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation tests in matched observational studies. *Biometrika*, *74*, 13–26.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika*, *75*, 577–581.

Rosenbaum, P. R. (1989). Sensitivity analysis for matched observational studies with many ordered treatments. *Scandinavian Journal of Statistics*, *16*, 227–236.

Rosenbaum, P. R. (1991a). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*, 901–905.

Rosenbaum, P. R. (1991b). Sensitivity analysis for matched case-control studies. *Biometrics*, *47*(1), 87–100.

Rosenbaum, P. R., & Krieger, A. M. (1990). Sensitivity of two-sample permutation inferences in observational studies. *Journal of the American Statistical Association*, *85*(410), 493–498.

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

Rosenbaum, P. R., & Rubin, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B (Methodological)*, *45*(2), 212–218.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, *2*, 808–840.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Walsh, M. E., & Brabeck, M. M. (2006). Resilience and risk in learning: Complex interactions and comprehensive interventions. In R. J. Sternberg & R. F. Subotnik (Eds.), *Optimizing student success in school with the other three Rs: Reasoning, resilience, and responsibility* (pp. 113–142). Greenwich, CT: Information Age.

Walsh, M. E., Madaus, G. F., Raczek, A. E., Dearing, E., Foley, C., An, C., Lee-St. John, T. J., & Beaton, A. (2014). A new model for student support in high-poverty urban elementary schools: Effects on elementary and middle school academic outcomes. *American Educational Research Journal*, *51*, 704–737.

What Works Clearinghouse. (2011). *Procedures and standards handbook*. Retrieved February 23, 2012, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf