Educational Evaluation and Policy Analysis
March 2014, Vol. 36, No. 1, pp. 44–66
DOI: 10.3102/0162373713485814

Assessing the Use of Aggregate Data in the Evaluation of School-Based Interventions: Implications for Evaluation Research and State Policy Regarding Public-Use Data

Robin T. Jacob
University of Michigan

Roger D. Goddard
Mid-Continent Research for Education and Learning

Eun Sook Kim
University of South Florida

It is often difficult and costly to obtain individual-level student achievement data, yet, researchers are frequently reluctant to use school-level achievement data that are widely available from state websites. We argue that public-use aggregate school-level achievement data are, in fact, sufficient to address a wide range of evaluation questions and the use of this data is more appropriate than commonly thought. Specifically, we explore (a) when point estimates and standard errors differ between models that use individual student-level data and those that use aggregate school-level data, (b) the potential for conducting subgroup and nonexperimental analyses with aggregate data, and (c) the metrics that are currently available in state public-use data sets and the implications these have for analyses.

Keywords: *school-level data, multilevel modeling, educational evaluation*

Increasingly, educational evaluations that rigorously assess the impact of school-based programs rely on state or district assessment data to test for effects on student achievement. Researchers rely on such extant data more frequently because (a) the cost of administering assessments to students is high, (b) the burden of additional testing is often a disincentive to schools and teachers to participate in research, and (c) No Child Left Behind has heightened interest in the outcomes of educational interventions on state assessments at all levels of the policy environment. Because schools operate in complex organizational and social environments with multiple levels of influence (e.g., students, classrooms, schools, families, and districts), in an ideal world, evaluators would always have access to individual-level student achievement and sociodemographic data to test the impact of an educational intervention.

Individual-level data provide researchers with the greatest flexibility in conducting analyses by enabling them to follow individual students over time, conduct subgroup analyses based on the characteristics of individual students, and by enabling the modeling of slopes in tests of cross-level interactions (e.g., whether strong school leadership attenuates the socioeconomic status–achievement relationship).

Unfortunately, the reality is that it is often difficult and costly to obtain student-level data, especially when one needs to obtain data from many districts within a single state. Indeed, privacy concerns and resource constraints have made state officials more reluctant to make student-level data available to researchers. However, school-level achievement data by grade and subject are widely available and easily downloadable from state department of education websites. The combination of these factors suggests that researchers are increasingly likely to use aggregate school-level achievement and demographic data in assessments of the relationships between school-level interventions and student achievement.

In this article, we argue that public-use aggregate school-level achievement data often are, in fact, sufficient to address a wide range of policy relevant issues related to school-based interventions and that the use of aggregate data is often more appropriate than commonly thought. In making this argument, we explore the adequacy of aggregate school-level data for evaluating the impact of a school-level intervention in the context of a randomized trial and in nonexperimental studies. Specifically, we explore whether point estimates and standard errors differ between models that use individual student-level data (often estimated using hierarchical linear models [HLM] or other "multilevel" models) and those that use aggregate school-level data. To do so, we compare under what circumstances one will obtain the same point estimates and standard errors when estimating the following two models:

1. A multilevel analysis of student-level achievement data:

Level 1:

$$Y_{ij} = \gamma_{0j} + \left(\sum_{z>0} \gamma_{zj} X_{zij} \right) + \varepsilon_{ij},$$

where Y_{ij} is the student-level achievement score for student i in school j, γ_{0j} is the regression-adjusted mean value of student achievement for school j, X_{zij} is the value of the zth student-level covariate for student i from school j, and ε_{ij} denotes the residual for student i from school j, which is assumed to be independently and identically distributed.

Level 2:

$$\gamma_{0j} = \beta_0 + \beta_1 T_j + \left(\sum_{s>0} \beta_s Z_{sj} \right) + \mu_j,$$

where β_0 is the grand mean of the regression-adjusted achievement score for the average school, T_j is an indicator variable equal to one if school j is in the treatment group or zero otherwise, β_1 represents the relationship between treatment status and achievement, Z_{sj} is the sth school-level covariate for school j, β_s represents the relationship between the sth school-level covariate and student achievement, and ε_j is the residual error for school j, which is assumed to be independently and identically distributed.

2. An (unweighted) ordinary least squares (OLS) model of aggregate school-level achievement:[1]

$$Y_j = \beta_0 + \beta_1 T_j + \left(\sum_{s>0} \beta_s X_{sj} \right) + \varepsilon_j,$$

where Y_j is the average achievement score for school j at a particular grade level, β_0 is the intercept, T_j is an indicator variable equal to one if school j is in the treatment group or zero otherwise, β_1 is the relationship between treatment status and achievement, X_{sj} is the value of the sth school-level covariate for school j, β_s is the relationship between the sth school-level covariate and student achievement, and ε_j is the residual error for school j, which is assumed to be independently and identically distributed.

It has been previously established that under certain conditions the results obtained from the OLS analysis of school-level data will be identical to those obtained from the analysis of nested data in a multilevel framework. Specifically, if (a) school-level average achievement scores

include all of the student-level achievement scores (i.e., if no student-level scores are excluded from the school average), (b) the data are completely balanced (i.e., have the same number of students per school in all schools), and (c) no covariates are included in the model, then one will obtain identical point estimates and standard errors whether analyzing data using a two-level HLM (with students nested within schools) with student-level achievement scores as the outcome measure or whether using OLS with aggregate school-level achievement scores as the outcome measure (Raudenbush & Bryk, 2002).

However, such conditions rarely hold in practice. The number of students per school will almost always vary, unless a specific number of students are sampled from each school, and even then missing data are likely to result in unequal numbers of students per school. If the number of students per school is not the same across all schools in a sample, then the point estimates and standard errors of the two models will not be identical, because the HLM analysis takes the number of students per school into account and weights the data accordingly. Furthermore, when covariates are included in the model, the variance–covariance structure of the model will change so that even in the unlikely case of completely balanced data, the two models will not be identical.

The availability of covariates at the student level can also alter the results. Covariates serve two purposes in the context of a school-level intervention study. First, they can improve the precision of the estimate of program impact—this is their primary value in the context of cluster randomized control trials (RCTs). Second, they can help to reduce bias in nonexperimental studies when the treatment and control groups vary significantly on one or more observable characteristics. Thus, having access to a restricted-use student-level data file offers the possibility of including additional student-level covariates in the models, which might both improve the precision of the estimates and help to reduce any bias. Previous research (e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007) suggests that aggregate school-level pretests can be as effective in improving the precision of point estimates as are individual student-level pretests. More recent research by Konstantopoulos (2012) takes this one step further and shows that when cluster effects are high (i.e., the school-level intraclass correlation coefficient [ICC] is greater than 0.10—a common scenario when student test scores are the dependent variable) and the student-level covariates are group-mean centered, school-level covariates will do more to increase precision than student-level covariates. However, when clustering effects are smaller (i.e., ICCs are less that 0.10) and lower level covariates are grand-mean centered or uncentered, lower level covariates (e.g., student-level covariates) increase power more than top-level covariates.

Finally, the presence of heterogeneous treatment effects (at either the student or the school level), which are common in educational interventions, may have implications for the adequacy of aggregate school-level data. Mathematically, the same principles apply in the presence of either heterogeneous or homogeneous treatment effects—that is, in the case of a completely balanced design and no covariates, the two models will yield identical results. However, no one has previously explored how much the degree of unbalance and the inclusion of covariates affect the parameters of interest in the presence of heterogeneous treatment effects. Moreover, the presence of heterogeneous treatment effects has implications for subgroup analyses that the researcher may wish to undertake.

Thus, this article is primarily concerned with understanding how much difference these factors (i.e., imbalance in the number of students per school and the inclusion of covariates) make in common empirical settings, including situations in which treatment effects are heterogeneous at either the student or school level. The article will address the following questions related to modeling the impact of school-level treatments on student outcomes:

1. By how much do the point estimates and standard errors differ between multilevel analyses versus single (school) level analyses in real-world situations, in which the number of students per school varies, in which covariates are included in the models, and in which there are heterogeneous treatment effects?

2. What is the potential and what are the limitations for conducting analyses with aggregate data, particularly in situations in which there are heterogeneous treatment effects?
3. What is the current status and relative utility for school effects research of publicly available achievement data across the states?

We begin by exploring the implications of using aggregate data in random assignment studies. Using simulations, we explore the influence that unbalanced data and the inclusion of covariates have on estimates obtained from various model specifications and under various scenarios, including the presence of heterogeneous treatment effects and the inclusion of covariates. We then explore in practice, using empirical data, how substantially the point estimates from school-level and multilevel models differ when the data are not completely balanced and when covariates are included at the student and school levels. Next, we discuss the ways in which aggregate data can and cannot be used to explore heterogeneous treatment effects. We then consider under what conditions aggregate data are sufficient for conducting strong nonexperimental analyses. Finally, we conclude the article by describing the different metrics (e.g., scale scores, percent proficient) that are available in publicly available school-level aggregate data across states and discuss the implications these have for analyses.

Background and Literature Review

Using school-level data rather than individual-level data in regression analyses is an example of aggregation—in this case, aggregating the individual-level student achievement data to the school-level mean achievement score. Methodologists have been debating the adequacy of aggregate data for exploring relationships for many decades. A seminal article by Robinson (1950) showed that the correlation between two variables measured at the individual level and the correlation between those same two variables measured at the group level are not the same. For example, exploring the relationship between students' gender and their literacy achievement is not the same as exploring the relationship between the percentage of female students in a school and the average achievement of that school. This has been referred to as "the ecological fallacy," and has shaped much of the thinking about the use of aggregate data over the past half century, making many researchers wary of using such aggregate data.

In this article, we are not interested, as Robinson was, in the relationship between two individual characteristics, such as gender and achievement. We are interested in whether a group-level characteristic (in this case, whether the school participated in a particular intervention or not) is related to an individual-level characteristic—namely, achievement. Furthermore, we are concerned not with correlation coefficients but rather with point estimates from a regression. Kmenta (1971) demonstrated that when the value of each independent variable in a regression is the same for all members of the same aggregate unit (e.g., school), the point estimates from an aggregate model will be identical to those from an individual-level model. For example, if all individuals in a school have the same value for the treatment variable (because the school was the unit of random assignment in a random assignment study), an individual-level analysis or a group-level analysis will yield the same results. What Kmenta's analysis does not address is what happens in instances when individual- and group-level covariates are included in the model. It also fails to consider the precision of the estimates.

Another line of research that has shaped the thinking about the use of aggregate data involves the use of multilevel modeling. This work emphasizes the importance of taking clustering into account when analyzing nested data—for example, data in which students are nested within classrooms that are nested within schools—to ensure that the standard errors are estimated properly. Previous research has demonstrated that omitting a level of nesting will produce incorrect estimates of standard errors for *student-level independent variables*, thus leading to incorrect inferences (Moerbeek, 2004; Van den Noortgate, Opdenakker, & Onghena, 2005).

This line of research has led many to believe, incorrectly, that *any* analysis of student

achievement data that does not use multilevel modeling will distort standard errors and lead to incorrect inferences. This is not always the case. Here we explore whether one will obtain comparable point estimates and standard errors if one uses school-level aggregate achievement scores analyzed in an OLS framework to explore the relationship between a *school-level variable* and student achievement versus using student-level data in a multilevel analysis to explore the same relationship. Neither of these approaches is the same as analyzing *student-level achievement data* without accounting for the clustering of students within schools; a practice that *will* result in standard errors that are generally too small, and thus lead to incorrect inferences (Type I errors) about the relationships of the independent variables to the dependent variable.

Evaluating the Use of Aggregate Data in RCTs

We begin our discussion by exploring the adequacy of aggregate data in RCTs of school-based interventions. To do so, we run a series of simulations in which we explore a variety of scenarios and assess under what sets of circumstances models using aggregate data yield comparable results with those that use student-level data in a multilevel context. We then use actual data from a restricted-use data file to explore how the use of aggregate data may play out in real empirical settings, and further explore whether the inclusion of student-level covariates can have a large influence in reducing bias and increasing the precision of estimates.

Exploring Various Scenarios Using Simulated Data

To compare the estimates obtained from aggregate data and student-level data, we began with a simulated scenario in which we had a sample of 100 schools, with an average of 100 students per school. We then randomly assigned the number of students per school so that the distribution of students per school was normally distributed with a mean of 100 and a standard deviation of our choosing—if the standard deviation of the number of students per school was equal to 0, all schools would have 100 students per school. The higher the standard deviation in the number of students per school, the greater is the degree of imbalance in the number of students per school. We set the minimum number of students per school equal to 3.

Next, we randomly assigned approximately half of the schools to a treatment condition and half to a control condition. We began with a scenario in which treatment effects were heterogeneous at the student and school levels but random and normally distributed. To generate a simulated outcome measure, we randomly assigned a student-specific error term to each student that was normally distributed with a mean of 0 and a standard deviation of 0.1 (e_{ij}). We set the proportion of variance at the school level equal to 0.20 and calculated the resulting school-level error term for each student (v_j). We also assigned each student a pretest score with a mean of 0 and a standard deviation of 1.0 (pretest_{ij}). Finally, each student in the treatment group was assigned a heterogeneous treatment effect β, such that the average treatment was equal to .2. We allowed the treatment to vary randomly at the student level with a mean of 0 and a standard deviation of 0.06, and at the school level with a mean of 0 and a standard deviation equal to 0.05. In the end, each student was assigned a simulated outcome (y) equal to $(0.2 + \beta_{ij} + \beta_j) \times T + 0.7(\text{pretest}_{ij}) + v_j + e_{ij}$.

We then identified three different degrees of imbalance—one in which the standard deviation of the number of students per school was set equal to 0, so that all schools had exactly 100 students per school; one in which the standard deviation of the number of students per school was equal to 15 (moderately unbalanced), so that 95% of the schools would have between 70 and 130 students; and one in which the standard deviation of the number of students per school was set equal to 40 (highly unbalanced), so that 95% of the schools would have between 20 and 180 students.

Each simulated data set was replicated 500 times, and on each data set, we ran three separate models. The first is a school-level aggregate OLS model. The second is a weighted school-level aggregate OLS model, weighted by the number of students per school. We wanted to see whether weighting to account for variations in school size would improve the estimates. The

third is a multilevel HLM with students nested within schools. These three models represent different forms of fixed-effect or random-effect precision weighting. The HLM uses random-effects precision weights (which account for estimation error for each school and the variation in mean true outcomes across schools). The weighted aggregate OLS uses a form of fixed-effect precision weighting (which accounts only for estimation error for each school) and the simple aggregate OLS does not.

The average point estimates and standard errors from these three models and the three different levels of imbalance (constant, moderate, and high), for the 500 simulations, are shown in top third of Table 1. In the first column of the table (equal number of students per school), the point estimates and standard errors for Models 1, 2, and 3 are identical, as expected.

The second column illustrates what happens when there is a moderate degree of imbalance in the number of students per school. Again, looking at the point estimates and standard errors for the first three models, we see that there is little difference among them. Finally, the third column represents a scenario in which the number of students per school is highly unbalanced. Even under this scenario, there is little difference among the point estimates and standard errors of the three models.

Table 1 also displays the frequency with which the impact estimates and standard errors from Models 1 and 2 differ from the same estimates obtained from Model 3 (the multilevel model) by more than 10%. The estimated impact from Model 3 is approximately 0.20, so a 10% deviation is equivalent to a difference in estimated impact of 0.02. In our data, this is equal to approximately 0.03 standard deviations.[2] In the case of moderate imbalance, the difference between the point estimates and standard errors of Models 1 and 2 never differed from the estimates of Model 3 by more than 10%. When the number of students per school was highly imbalanced, the estimated impact between Models 1 and 3 differed by more than 10% in 4% of the cases and the standard errors differed by more than 10% in 1.2% of the cases. Differences greater than 10% occurred only in simulations where the minimum number of students per school was equal to 3, a situation that, as we discuss in more detail below, is unlikely to occur in practice. Furthermore, even in simulations where the minimum school enrollment dropped to 3, and the impact estimates differed by more than 10%, the absolute maximum difference, across all 500 simulations (shown in the final column of Table 1), never exceeded 0.05 standard deviations, making it unlikely that the substantive interpretation of the findings would differ based on the estimate method used.

When the aggregate data were weighted by the number of students per school (Model 2), the impact estimates from Models 2 and 3 differed by more than 10% in 6.6% of the cases and the standard errors differed by more than 10% in 6% of the cases. The differences among the estimates in these three models are due, in part, to the different weighting schemes used in estimation. Model 1, the school-level model, gives equal weight to each school. Model 3, the multilevel model, is a precision weighted model, which accounts for estimation error for each school and the variation in true mean outcomes across schools. We had hypothesized that weighting the school-level model by the number of students per school (Model 2) would yield impact estimates that were closer to the multilevel model, because the weighting would guard against small schools having an undue influence on the results. However, this hypothesis was not confirmed. Using the weighted OLS generally resulted in estimates that were further away from the impact estimates obtained from the multilevel model. Thus, we do not recommend weighting by the number of students per school.

In the second panel of Table 1, we explore the case of systematic heterogeneity of treatment effects. Panel 2 shows a scenario in which instead of allowing the treatment to vary randomly across students and schools, we constructed the treatment effect so that it varies based on a randomly generated pretest, such that students with a higher pretest had a smaller treatment effect. In this scenario, each student was assigned a simulated outcome (y) equal to $(0.2 - 0.05 \times \text{pretest}_{ij} + \beta_j) \times \text{T} + 0.7(\text{pretest}_{ij}) + v_j + e_{ij}$. The results from these simulations were consistent with the findings from the scenario in which the treatment effect varied randomly. The

TABLE 1
Average Estimated Impacts and Standard Errors of Simulated Data.

	Estimated impacts and standard errors			% of estimates that differed[a] by >10%	Absolute max. diff. across 500 simulations[b]	% of estimates that differed[a] by >10%	Absolute max. diff. across 500 simulations[b]
	Degree of imbalance						
	Constant	Moderate	High				
Estimation method	*SD* = 0	*SD* = 15	*SD* = 40	*SD* = 15	*SD* = 15	*SD* = 40	*SD* = 40
Treatment effect varies randomly across students and schools							
School-level OLS	.1984 (.0359)	.2004 (.0361)	.1989 (.0376)	0% (0%)	.0035 (.0001)	4.4% (1.2%)	.0347 (.0041)
Weighted school-level OLS	.1984 (.0359)	.2002 (.0360)	.1989 (.0359)	0% (0%)	.0146 (.0010)	6.6% (6.0%)	.0412 (.0100)
Student-level HLM	.1984 (.0359)	.2004 (.0361)	.1988 (.0365)				
Systematic heterogeneity in the treatment effect							
School-level OLS	.2032 (.0347)	.1995 (.0350)	.1973 (.0365)	0% (0%)	.0035 (.0001)	4.2% (1.2%)	.0224 (.0040)
Weighted school-level OLS	.2032 (.0347)	.1994 (.0350)	.1974 (.0350)	0% (0%)	.0136 (.0011)	7.6% (4.6%)	.0472 (.0035)
Student-level HLM	.2032 (.0347)	.1995 (.0350)	.1974 (.0355)				
Systematic heterogeneity in the treatment effect and school-level covariate							
School-level OLS	.2007 (.0325)	.2020 (.0324)	.1997 (.0330)	0% (0%)	.0033 (.0001)	0.4% (8.6%)	.0220 (.0051)
Weighted school-level OLS	.2007 (.0325)	.2020 (.0324)	.1993 (.0327)	0% (0%)	.0138 (.0012)	4.0% (8.4%)	.0332 (.0059)
Student-level HLM	.2007 (.0325)	.2020 (.0324)	.1996 (.0317)				

Note. OLS = ordinary least squares; HLM = hierarchical linear model; ICC = intraclass correlation coefficient. The information in this table is derived from 500 simulations. Estimated standard errors are shown in parentheses. *SD* indicates the standard deviation in the number of students per school. Simulations assumed a school-level ICC of 0.20. Assumes 100 schools with an average of 100 students per school.

[a]All differences relative to Model 3.

[b]Absolute maximum difference observed across all 500 simulations.

estimates between Models 1 and 3 only exceeded 10% in instances where the minimum number of students per school dropped to 3, and even among those simulations, the maximum difference in the impact estimates is only ever 0.03 standard deviations.

In the third panel of Table 1, we add a school-level covariate to the model used in Panel 2—the average school-level pretest. We wanted to see whether the presence of a covariate would alter our findings. It appears that in samples with a high degree of imbalance, adding a covariate that accounts for some of the heterogeneity in treatment effects helps to reduce the number of times that impact estimates from the two models diverge substantially from one another, but it increases the likelihood that the standard errors will be different. Nonetheless, even the largest differences would not cause the substantive interpretations of the findings to change.

Finally, we conducted simulations (available from the authors) in which we varied the number of schools in the sample (20, 30, 40, 60, or 100) and the intraclass correlations (.10, .15, and .20). We find that the number of times that the impact estimates from the various models differ by more than 10% from Model 3 increases as (a) the degree of imbalance in the number of students per school increases and (b) the total number of schools in the sample decreases. When there was only moderate variation in school size, the difference between Models 1 and 3 rarely exceeded 10% and never differed by more than 0.05 standard deviations, even with as few as 20 schools. In simulations with high variation and only 30 schools, the maximum difference between the impact estimates from Models 1 and 3 did not exceed 0.05 standard deviations as long as the minimum enrollment in any school did not drop below 5 students. With highly imbalanced enrollment numbers and only 20 schools in the sample, the minimum enrollment among the schools had to be at least 25 to ensure that the two methods always yielded results that were substantively similar. Yet even when the enrollment numbers dropped below 25, the estimates from the two models differed from one another substantively only on rare occasions. With 20 schools and high variation in enrollment, the difference in impact estimates between Models 1 and 3 exceeded 0.05 standard deviations in only 0.2% of the cases.

Thus, with our simulated data, it appears that under a wide variety of scenarios, including various degrees of imbalance in the number of students per school, the presence of heterogeneous treatment effects (either systematic or random), and the inclusion of school-level covariates, the two models (school-level OLS or student-level HLM) yield fairly comparable results. Researchers may wish to exercise caution in instances where the number of schools in the sample is small (30 or less), and there are schools in the sample with very small enrollment numbers (e.g., less than 5).

Comparing Point Estimates and Standard Errors Using Empirical Data

To illustrate how the use of aggregate school-level data plays out in practice, and to highlight some of the factors that researchers should consider when using aggregate data, we turn briefly to an analysis that uses empirical (as opposed to simulated) data. The data used for these analyses were originally collected as part of a study designed to explore the relationship between school leadership, climate, and teacher practice on student achievement in Michigan elementary schools. The participants for that study were selected from a stratified random sample of Michigan's noncharter public elementary schools serving at least Grades 4 and 5 in 2004–2005. The final analytic sample includes test scores from 5,031 students in 78 schools. Student- and school-level descriptive statistics for this sample are reported in Table 2.

Student-level achievement data were obtained from the Michigan Department of Education. The outcome measures were student scaled scores on the 2005 Michigan Educational Assessment Program (MEAP) fourth-grade mathematics and reading tests. For each student, measures of ethnicity, gender, special education program status (yes/no), limited English proficient (LEP) status (yes/no), and free or reduced price lunch status (yes/no) were also obtained.

To obtain school-level data, we downloaded data on school size, percentage of minority students, percentage of students eligible for free or reduced price lunch, prior achievement

TABLE 2
Descriptive Statistics by Condition (n = 78 Schools).

	Control		Treatment	
Covariate	*M*	*SD*	*M*	*SD*
School level				
Urban	0.76	0.43	0.75	0.43
Size	370.83	113.24	371.53	109.88
% Minority	0.22	0.29	0.20	0.27
% Free lunch	0.37	0.26	0.36	0.25
Prior math	428.14	11.02	428.54	10.49
Prior reading	427.11	11.91	427.78	11.24
% Spec Ed	0.11	0.06	0.11	0.06
% LEP	0.03	0.13	0.05	0.13
% Female	0.50	0.07	0.50	0.07
Student level				
Spec Ed	0.11	0.31	0.11	0.31
LEP	0.04	0.19	0.04	0.19
Free lunch	0.34	0.47	0.34	0.47
Minority	0.20	0.40	0.18	0.38
Female	0.50	0.50	0.50	0.50

Note. MEAP = Michigan Educational Assessment Program; Urban = whether the school was in a metropolitan statistical area or not; Size = total school size; Minority = non-White student; Free lunch = free or reduced price lunch student; Prior math = average prior MEAP math achievement; Prior reading = average prior MEAP reading achievement; Spec Ed = special education student; LEP = limited English proficient student. All variables are from the original data of this study. However, the control and treatment groups are randomly generated 100 times and the mean and standard deviation are the average across 100 treatment control assignments.

(measured by the prior year's MEAP 4th-grade pass rates), and urbanicity (whether the schools were classified in a metropolitan statistical area) from the Michigan Department of Education website. School-level aggregate achievement scores were constructed by averaging the individual student MEAP scores contained in the student-level file described above. For comparative purposes, we also downloaded school MEAP percent proficient for fourth-grade mathematics and reading in 2004 and 2005. In 2004–2005, percent proficient scores were the only scores available for download from the state website.

As these data were not part of an evaluation study, we used a random number generator to assign schools to a simulated treatment or control condition. Specifically, each school was assigned a random number and schools with numbers over the median were assigned a value of one and those below the median were assigned a value of zero. We then simulated a situation in which the treatment made a greater impact on low-achieving students but had less impact on high-achieving students. In our data set, student math scores ranged from 439 to 743, with a mean of 549 and a standard deviation of 29. For the students at the extreme low end of the distribution (i.e., with scores below 480), we simulated an average treatment effect of 13 points, and for the students with scale scores between 480 and 520, we simulated an average treatment effect of 10 points. For all others, we simulated an average treatment effect of 7 points such that the overall average treatment effect in the sample was approximately 10 points. The student-level standard deviation for the MEAP is equal to 29, so a 10-point treatment effect is equivalent to approximately one third of a standard deviation.

Establishing the Equivalency of the Data. One factor that researchers need to consider when using aggregate data to address questions of program impact is whether the publicly available school-level average achievement scores (provided by states) and the average of the individual-level student achievement scores (calculated from the student-level data) are, in

fact, identical. Although school-level average achievement data are available in many states, they may not be based on exactly the same data as one would find in a restricted-use student-level file. Because state data reporting requirements often restrict the reporting of data for groups of fewer than 10, it is possible that some students who would be included in a restricted-use student-level data set would be excluded from a publicly available school-level data set compiled by the state.

In Michigan in 2009–2010, there were 38 elementary schools with at least one but fewer than 10 third-grade students—data for third-grade students from these schools would not be available in the public-use files. This represents approximately 2% of all elementary schools in Michigan. Nine of these schools (24%) are schoolwide Title 1 schools as opposed to 46% of schools statewide, and 31 of these schools (81.6%) are rural schools as opposed to 24% of schools statewide. This suggests that the schools that are being excluded could systematically introduce bias into results based only on publicly reported data. Thus, if more complete demographic data can be obtained from restricted-use student-level data sets, then there may be some advantage to these data over publicly available school-level mean achievement data.

Therefore, we began our empirical analysis by establishing that the two data sources are equivalent in our data. In our data, the publicly available school-level average achievement scores (provided by state) and the average of the individual-level student achievement scores (calculated from the student-level data) are, in fact, identical. All of the 78 schools in our student-level file were also included in the school-level data, and the values for the publicly available school-level variables exactly matched the values obtained from averaging the corresponding variables in the restricted-use student-level file. However, had we selected a different sample of schools, this might not have been the case. Furthermore, different states have different reporting requirements. The appendix lists the reporting requirements for all 50 states, along with the data available from those states. Notably, there is a considerable range in the minimum number of students required for reporting among the various states; some states report results for as few as 5 students while others report only for groups of 40 or more. Researchers should carefully explore whether certain schools might be excluded from public-use school-level files when considering the use of publicly available aggregate data, because data on smaller schools in rural locations could be systematically excluded from the data. This may be of particular concern for states with high minimum reporting requirements or for states with many small schools.

Comparing Various Estimation Techniques Using Empirical Data. Next, we explored whether the results we demonstrated using simulations could be replicated using the empirical data, with real covariates and with variation in the number of students per school that ranged from 17 to 224. We estimated two models. In the first, we used student-level scaled scores on the MEAP as the outcome variable and estimated a two-level HLM with the generated treatment variable as the only predictor variable at Level 2. In the second, we used school-level average scale scores on the MEAP (constructed by averaging individual fourth-grade scores in each school) as the outcome variable and estimated an unweighted OLS model with the generated treatment variable as the only predictor. The results are shown in Table 3 (Models 1 and 4). As predicted, because the data are not completely balanced, the point estimates and standard errors for the OLS and HLM estimates are not identical, however, they are quite comparable. For math, the coefficient on the treatment variable in the OLS model was 10.244 with a standard error of 2.630 while the coefficient for the treatment variable from the HLM was 10.082 with a standard error of 2.541. The findings for reading (not shown here) were similar.

We also ran a set of models in which we included school-level covariates to see whether we would obtain the same results with the HLM and OLS model. In these models, we included the following variables: whether the school was in a metropolitan statistical area, school size, percent minority, percent free or reduced price lunch, and the school average prior math achievement. School average achievement scores were constructed by averaging the individual-level fourth-grade scores in each school. As this is a simulated random assignment study, with the addition of covariates, we would expect

TABLE 3
Comparison of Models That Include Different Sets of Covariates for Estimating Simulated Impacts on Math Achievement.

	OLS[a]			HLM[b]			
Covariate	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
School level							
Urban		2.768 (2.668)	1.195 (2.344)		2.542 (2.672)	2.944 (2.717)	0.452 (2.438)
Size		−0.938 (1.070)	−2.766* (0.987)		−0.893 (1.115)	−0.746 (1.136)	−2.209* (1.045)
% Minority		−2.852* (1.336)	−1.125 (1.275)		−2.806* (1.398)	0.233 (1.495)	2.370 (1.410)
% Free Lunch		−1.314 (1.348)	−3.990* (1.329)		−1.363 (1.392)	0.989 (1.446)	−2.046 (1.423)
Prior math		5.559* (1.153)	6.046* (1.008)		5.782* (1.231)	5.814* (1.246)	5.903* (1.106)
% Spec Ed		—	−29.378 (1.881)			—	−4.280 (16.593)
% LEP		—	26.847* (6.583)			—	37.892* (7.686)
% Female		—	−29.929* (13.264)			—	−23.795 (13.821)
Treatment	10.244* (2.630)	9.805* (2.079)	9.892* (1.831)	10.082* (2.541)	9.821* (2.081)	9.825* (2.114)	9.861* (1.884)
Student level							
Spec Ed		—	—			−18.089* (1.288)	−18.088* (1.290)
LEP		—	—			−4.565 (2.593)	−8.437* (2.732)
Free lunch		—	—			−8.469* (0.950)	−8.411* (0.949)
Minority		—	—			−11.090* (1.335)	−10.844* (1.336)
Female		—	—			−2.859* (0.782)	−2.819* (0.782)
MDE (MDES)	7.364 (0.253)	5.821 (0.201)	5.127 (0.177)	7.115 (0.245)	5.827 (0.201)	5.919 (0.204)	5.275 (0.182)

Note. OLS = ordinary least squares; HLM = hierarchical linear model; MEAP = Michigan Educational Assessment Program; Urban = whether the school was in metropolitan statistical area or not; Size = total school size; Minority = non-White student; Free lunch = free or reduced price lunch student; Prior math = average prior MEAP math achievement; Spec Ed = special education student; LEP = limited English proficient student; MDE = minimum detectable effect; MDES = minimum detectable effect size; ICC = intraclass correlation coefficient. The results are the average of 100 replications with simulated treatment effects added to the original data. Standard errors are shown in parentheses. Models 1 and 4 only include a treatment variable. ICC math = .142.

[a]Model includes 78 schools.

[b]Model includes 5,031 students in 78 schools.

*$p \leq .05$.

the point estimates to remain relatively stable and the standard errors to decrease, but would expect the OLS model and HLM to yield comparable results.

The results of the models with school-level covariates included are shown in Table 3 (Models 2 and 5). With the addition of the covariates, the standard error on the treatment variable is reduced as expected and the coefficient for the treatment variable is also changed slightly, but the HLM and OLS model continue to yield very comparable results. The coefficient for the treatment variable in the OLS model is 9.805 with a standard error of 2.079. In the HLM, the coefficient for the treatment variable is 9.821 with a standard error of 2.081. Thus, even with imbalanced real empirical data and the inclusion of covariates at the school level, we obtain quite comparable results whether we use aggregate school-level data or individual student-level data. With the point estimate for the OLS model exceeding the value of the estimate obtained from the HLM analysis by less than 2% and both estimates attaining statistical significance, it is quite unlikely that the substantive theoretical and policy interpretations of these results would differ based on the use of student- versus school-level data.

Inclusion of Student-Level Covariates. Although we have established that the impact estimates obtained from an aggregate school-level file and a student-level restricted-use file will, under a wide variety of circumstances, yield quite comparable results, one potential advantage of obtaining student-level data from a restricted-use data file is that one may be able to obtain student-level demographic or other covariate data not available in a public-use school-level file. Including these student-level covariates could potentially increase the power of a study beyond that which could be achieved with school-level covariates alone.

In our empirical data, we have three variables at the student level that were not available at the school level: special education status (yes/no), LEP status (yes/no), and gender, and two variables at the individual level that were also available at the school level in the public-use data: minority status and free or reduced price lunch status. Using our empirical data, we ran another set of models to explore differences across three different analytic scenarios. Our findings are reported in Table 3. Model 2 reports results for an OLS model that includes only the school-level covariates that were publicly available. Model 6 represents an HLM in which we retained the school-level covariates publicly available for OLS Model 2 and to which we added the student-level covariates that were available in our restricted-use data file at Level 1. Model 2 represents the best one could do with the public-use data whereas Model 6 represents the addition of student-level variables provided in a restricted-use data set. Models 3 and 7 explore whether the precision of our estimates would increase if we included aggregate versions of these additional student-level covariates at the school level. We used individual-level data to create school averages for special education, LEP, and gender, and added these variables to Model 7 at Level 2 and retained the individual-level covariates at Level 1. Model 3 includes the covariates from Model 2 and also the student-level variables that were available in the restricted-use file aggregated to the school level. The scenario represented in Model 3 indicates what might happen if states were to make a wider range of aggregate data available to researchers.

To provide a sense of how meaningful the differences in standard errors in the various models are, we calculated the minimum detectable effect (MDE) for the OLS model and HLM shown in Table 3. The MDE is the smallest true impact you are likely to be able to detect with your model (Bloom, 1995). With alpha = .05, a two-tailed test, and power equal to .80, the MDE = 2.8 × (standard error of the estimate).

One can also obtain a minimum detectable effect size (MDES) by dividing the MDE by the standard deviation of the outcome. The standard deviation of the MEAP in our data is 29 points, so a 5-point difference is equivalent to an effect size of 0.17 standard deviations. The final panel in Table 3 shows the MDEs and MDESs for each of the models.

The first thing to notice is the precision of the models that include covariates is quite comparable. All five models with covariates yield a MDES between .18 and .20 (the unconditional

model shown in Model 1 had a MDES of around .25 by comparison). The second thing to notice is that adding student-level covariates to the HLM (Model 6) did little to improve the precision of our estimates. In fact, adding the student-level covariates actually *increased* the MDE from what was obtained with OLS Model 1—likely because the student-level covariates are somewhat redundant with the school-level covariates, and thus they are reducing the degrees of freedom in the model without adding any explanatory power. The MDE for OLS Model 2 is 5.821 points and for HLM Model 6 is 5.919 points. Although we were not able to test whether this finding also holds for individual pretest data, because we were not able to obtain individual pretest data from the state, our findings are consistent with that of Bloom et al. (2007), who found that aggregate school-level pretests were as effective in increasing the precision of a randomized study as were individual student-level pretests. Our findings are also consistent with Konstantopoulos (2012), who showed that school-level covariates can do more to increase the power than student-level covariates, when the ICC is greater than 0.10 and data are group-mean centered. The ICC in our data is 0.14, although the student-level covariates are uncentered.

However, the aggregate student-level variables that are included in Model 7 do appear to decrease the MDE somewhat from 5.821 points in OLS Model 2 to 5.2752 points in the HLM Model 7—the differences translate to a difference in effect size of around 0.02 standard deviations. Thus, one potential benefit of obtaining a student-level restricted-use data file may be that it enables the creation of additional school-level variables, which can increase the proportion of between-school variation that is explained in the model and thereby increase the power of the study. In this case, the increase in power gained by adding these school-level covariates is relatively small, but in other circumstances, it might be more sizable. Finally, OLS Model 3 shows what would happen if the student-level variables in the restricted-use file were available in aggregate form to researchers. The MDE for OLS Model 3 is slightly smaller even than what one would get using the restricted-use student-level data as shown in Model 7, arguing for the inclusion of a greater set of aggregate variables in public-use files.

Exploring Heterogeneity in Outcomes: Subgroup Analyses

When assessing the overall impact of a school-level intervention, our analyses indicate that under most circumstances aggregate data work quite well. However, in addition to understanding overall program impacts, researchers are also often interested in understanding the heterogeneity of treatment effects. This is often accomplished by conducting subgroup analyses. Subgroup analyses can be somewhat more difficult to undertake with aggregate data. For example, if you want to know the impact of a particular educational program on the lowest achieving students in your sample, without individual-level data, one generally could not select the subgroup of interest for analyses. However, No Child Left Behind requires reporting of results disaggregated by subgroup, which could potentially be employed to answer some questions about the heterogeneity of effects. Currently, states are required to report disaggregated results for the following subgroups: (a) students who are in ethnic minorities, (b) students who speak English as a second language, (c) students who are economically disadvantaged, and (d) students who are emotionally, physically, or mentally disabled to the extent that they need Individualized Education Plans (IEPs). Using these disaggregated results, one can conduct analyses to estimate subgroup differences. Instead of using individual-level data to select subgroups, publicly reported average school and grade-level subgroup scores can be used as outcome measures. As we demonstrate below, as long as the number of students per subgroup per school is greater than five in most schools, conducting analyses in this way will yield comparable results with what would have been obtained using individual student-level data.

Table 4 shows simulations in which we used (a) individual student data and selected students who scored in the bottom quartile of

TABLE 4
Average Estimated Impacts and Standard Errors of Subgroups Based on Simulated Data.

	Estimated impacts and standard errors		% of estimates that differed[a] by >10%	Absolute max. diff. across 500 simulations[b]	% of estimates that differed[a] by >10%	Absolute max. diff. across 500 simulations[b]
	Degree of imbalance					
	Moderate	High				
Estimation method	*SD* = 15	*SD* = 40	*SD* = 15	*SD* = 15	*SD* = 40	*SD* = 40
School-level OLS	.2649	.2659	0%	.0081	1%	.0352
	(.0369)	(.0389)	0%	(.0003)	1.2%	(.0052)
Weighted school-level	.2649	.2642	0%	.0134	1.4%	.0331
OLS	(.0367)	(.0370)	0%	(.0013)	6.0%	(.0046)
Student-level HLM	.2648	.2650				
	(.0368)	(.0377)				

Note. OLS = ordinary least squares; HLM = hierarchical linear model; ICC = intraclass correlation coefficient. The information in this table is derived from 500 simulations. Estimated standard errors are shown in parentheses. *SD* indicates the standard deviation in the number of students per school. The *SD* in the number of students per school reflects the initial imbalance in the simulated schools where there were on average 100 students per school. Simulations assumed a school-level ICC of 0.20. Assumes 100 schools with an average of 100 students per school.

[a]All differences relative to Model 3.

[b]Absolute maximum difference observed across all 500 simulations.

the distribution on the pretest (row labeled student-level HLM) and (b) the school-level average test scores of the students who scored in the bottom quartile on the pretest to estimate impacts (rows labeled school-level OLS and weighted school-level OLS). The school-level averages were created by selecting the students in each school who scored below the 25th percentile and then averaging their scores. The results show that you will get comparable results, whether you estimate subgroup impacts using individual student data or aggregate subgroup data, even when the initial imbalance in number of students per school is relatively high.

Although, for ease of explanation, we use prior achievement as the example in the previous simulation, unfortunately data available from states are not typically disaggregated by prior achievement levels. We therefore used our empirical data, and conducted analyses that used minority status as the subgroup of interest. Again, we ran a two-level model selecting individual students with minority status and then ran an OLS model in which the dependent variable was the average score for the minority students in the school. The OLS model had a treatment effect of 10.688 with a standard error of 3.584, and the HLM yielded a treatment effect of 11.031 with a standard error of 2.872—although not identical, the two point estimates differ only by 0.01 of a standard deviation (recall that the standard deviation of the outcome is equal to 29) and support similar substantive interpretations. The HLM is somewhat more precise, with a MDES of .27 as opposed to .33 for the OLS model. In our sample, 40 of the 78 schools in the original sample had fewer than five minority students per school (51% of the sample); 5 schools had no minority students and 10 schools had only one.

Unfortunately, a large downside of using data disaggregated by subgroup to explore heterogeneity of effects is that states are not required to report disaggregated results for schools that do not have enough students in a particular subgroup to protect the privacy of the students. As already noted, and shown in the appendix, each state can determine its own threshold for reporting. In most states, there must be at least 10 students per grade per subgroup for the state to report results. This means that many schools will not have disaggregated results available for subgroups of interest. In Michigan in 2009–2010, 580 of the 1,806 elementary schools in the state have between 2 and 9 African American third-grade students and thus no subgroup scores are reported for African Americans in

these schools. This represents 32% of the schools in the state.

In some instances, this problem can be ameliorated by looking at test scores for student subgroups across the entire school, rather than analyzing subgroup scores separately by grade, because many schools will meet minimum reporting numbers on a schoolwide basis even if they do not have sufficient numbers in any particular grade. However, for some research questions, such analyses may be inappropriate and some states may not make data aggregated in this way available.

One is also limited in the types of subgroup analyses that can be undertaken. The fact that data available from states are not typically disaggregated by prior achievement levels makes it difficult to explore heterogeneity of treatment effects based on student-level prior achievement. Unless there is a very high degree of homogeneity in terms of prior achievement within schools (a situation that is unlikely to occur in practice), schools in the bottom of the distribution will not be a good proxy for students in the bottom of the distribution. This is a clear limitation of having aggregate data, although it is less a limitation of aggregate data per se than of the data that states and districts currently make available. Furthermore, one would not be able to conduct subgroup analysis based on multiple characteristics simultaneously—for example, the impact of the program on students who are female and poor, which may be of interest in some circumstances.

However, one can obtain information about the heterogeneity of treatment effects based on school-level characteristics, including prior achievement. Using school-level data, researchers can explore whether the program was more effective for schools based on their size, location (rural vs. urban), grade configuration, racial/ethnic composition, and prior achievement, among other things. In the case of schoolwide interventions, such heterogeneity is often a key question of interest.

In general then, with aggregate data that are currently publicly available from many states, researchers can explore (a) whether the program being evaluated works better in certain types of schools than in others (e.g., low-achieving schools, schools in disadvantaged areas, schools with high proportions of minority enrollment, and small schools) and (b) under some circumstances, whether the program works better for some subgroups of students than others (e.g., ethnic minorities, students who speak English as a second language, students from disadvantaged backgrounds, and students with disabilities). However, researchers currently cannot explore whether the program works better for students with different levels of prior achievement (low-achieving students in high-achieving schools, for example), or whether the program works better for students based on multiple background characteristics. Finally, researchers will be limited in the types of subgroup analyses that they can conduct due to minimum reporting requirements of states, which typically require at least 10 students per group per grade for reporting. In cases where such data are available, researchers should use caution if the number of schools in the sample is small and the minimum enrollment in any school drops below 5.

The Use of Aggregate Data in Nonexperimental Analyses

Thus far, we have been focusing on the use of random assignment studies to evaluate the effectiveness of school-level interventions. However, under a variety of circumstances, random assignment may not be possible. Here we discuss the use of aggregate achievement data in nonexperimental studies and argue that in many circumstances aggregate data should be sufficient to conduct strong nonexperimental evaluations of school-level interventions as well.

First, we note that many of the most rigorous nonexperimental methods used for estimating program impacts, including interrupted time-series analysis and regression-discontinuity analysis, do not require the use of individual-level data for their basic estimation. The validity of a regression-discontinuity analysis, for example, depends primarily on identifying the correct functional form of the rating variable used to determine who will receive the treatment. For school-level interventions, the rating variable will always be a school-level variable and thus school-level aggregate data will be sufficient. As with RCTs, other covariates are

included in regression-discontinuity analyses primarily to increase the precision of the estimates, and thus an analogous argument to the one made regarding RCTs can be made about the adequacy of aggregate data in the context of a regression-discontinuity design.

Similarly, interrupted time-series analysis explores achievement trends over time and tries to estimate the impact of a school-level program by determining whether achievement differs after program implementation relative to a school's preprogram trend. The primary requirement for interrupted time-series analysis is that you have enough data points to establish pre- and post-trend lines, and in the case of a school-level intervention, the trend being explored is the trend in school-level average achievement scores over time. Again, the use of individual-level data is not necessary. As we have shown, an HLM with individual student achievement data will yield similar results as an OLS model with aggregate data, as long as the value of each of the independent variables in the model is the same for all members of the same aggregate unit (students in a given school in a given year). The same concept applies to the basic interrupted time-series analyses as well, because the independent variables in these analyses are treatment-by-year indicators, which will have the same value for all students within a school within a given year. See Bloom (1984) for an excellent example of the way in which aggregate data (as opposed to individual data) can be used in this way to estimate the impacts of a program.

However, to improve the quality of a nonexperimental design, like interrupted time-series, researchers will often add a set of matched comparison schools to the model. The question then becomes whether appropriate matches can be made using aggregate data. Researchers often use propensity scores to identify such schools. A general rule of thumb in identifying a good match suggests that the more covariates you include in the propensity score model, the better off you are, whether or not the additional covariates are directly related to the outcome (Hill, Reiter, & Zanutto, 2004; Rubin & Thomas, 1996). Covariates at the individual and aggregate levels have been shown to contribute to better matches (Gao & Fraser, 2010). To the extent that publicly available aggregate data have a limited set of covariates on which to match, using these data may put constraints on the analysis and result in suboptimal matches.

Yet, two recent syntheses of nonexperimental approaches to estimating program impact have concluded that there are three key conditions that need to be in place for nonexperimental analyses to yield findings that are comparable with RCTs—(a) the comparison group must be chosen from a sample with similar motivation and incentive to participate in the program (e.g., schools that all applied to participate in a program), (b) the comparison group must be located in close geographic proximity to the treatment group (e.g., in the same district), and (c) pretest scores must be available for the outcome of interest (Cook, Shadish, & Wong, 2008; Glazerman, Levy, & Myers, 2003). In the case of a school-level intervention, none of these three conditions requires the use of individual student-level data.

Furthermore, the most important variables on which to match in the case of a school-level intervention are school-level variables. This idea is supported by work in progress by Somers, Zhu, Jacob, and Bloom (2012), which shows that in the case of comparative interrupted time-series (CITS) analyses designed to assess the impact of a school-level intervention, the key variables on which to match are school-level test scores, and that including other variables, such as demographic data can actually reduce the adequacy of the match. As such, aggregate school-level data will be more than adequate to conduct rigorous quasiexperimental analyses in the context of a CITS analysis.

In fact, it is difficult to construct a scenario in which individual-level data are strictly necessary to achieve a good school-level match. One benefit of having individual-level prior achievement data would be to include information about the distribution of student achievement within schools rather than simply matching on the average level of prior achievement. However, this would not require the use of individual-level data per se. In fact, the percent proficient metrics that many states currently report could be used as matching vari-

ables to account for the variability in the distribution of students—for example, one could match based on average prior achievement and on the percentage of students in the school who fell below the "below basic" cutoff, who fell below the "basic" cutoff, and so on. Similarly, although restricted-use student-level data sets may provide a richer set of covariates on which to match, if these same covariates were available at the school level, they would likely be sufficient to achieve good matches.

In general, the various nonexperimental estimation methods lend themselves to the use of aggregate data. Although it is possible that better estimates of program impact could be obtained using student-level data, previous research also suggests that aggregate data are likely to be sufficient to provide estimates of program impact that are comparable with those achieved in RCTs.

Metrics Available

Up until this point, we have said nothing about the test score metrics that are publicly available in aggregate school-level files. The final consideration in determining the adequacy of school-level data for evaluating the impact of a program implemented at the school level is whether the metric in which the state reports achievement scores limits the analysis in any way. The No Child Left Behind Act only mandates that states report the percentage of students reaching proficiency standards, and as a result, many states only report percent proficient scores (i.e., the percentage of students receiving scores above a specified cut point) and not scaled scores in their aggregate public-use data files. A number of researchers have demonstrated that percent proficient metrics have poor statistical properties and can lead to distorted interpretations of achievement gaps and achievement trends (e.g., Ho, 2008; Holland, 2002). For example, the choice of a cut score can lead to inflated or deflated estimates of program impact depending on where along the distribution the cut score is located. Furthermore, changes can occur in the upper or lower parts of the distribution that will not be detected by the percent proficient metric if not enough students cross the proficiency threshold. For example, distributions with right censoring (e.g., 100% correct on a state test) can be seriously distorted in proficiency-based analyses. We therefore explore the implications of having only the percent proficient metric available to make inferences about achievement and assess which states have publicly available scores in these metrics.

To demonstrate the weakness of proficiency scores, we again simulated a treatment effect in our data, using a method similar to the one described above. Using the empirical data for the state of Michigan described earlier, we again simulated a situation in which the treatment made a greater impact on low-achieving students but had less impact on the students around the cutoff. For the students at the extreme low end of the distribution (i.e., with scores below 480), we simulated an average treatment effect of 12 points, and for the students with scaled scores between 480 and 520, we simulated an average treatment effect of 10 points. For all others, we simulated an average treatment effect of less than 7 points such that the overall average treatment effect in the sample was approximately 7 points. We then created proficiency scores based on these simulated scores and the designated proficiency level for the state of Michigan, which was 529. The proficiency scores are dichotomous variables—students either reached the proficient level or they did not. We then ran analyses using the two different outcome measures.

The results are shown in Table 5. The top half of the panel shows outcomes using scaled scores with a generated treatment variable and a simulated impact of approximately 7 points. The treatment effect is statistically significant in all four models. The bottom half of the panel shows the same analysis using percent proficient data. None of the estimated impacts are statistically significant and both the models with and without covariates indicate that the effect was near zero. A 7-point impact translates to an effect size of .23 in our data. This is a fairly substantial impact and yet, with the percent proficient metric, our simulated treatment appears to have no effect on achievement. This argues for using scaled scores or other

TABLE 5

Comparison of Scaled Score and Percent Proficient Metrics in Math Achievement: Coefficients and Standard Errors.

School-level covariates	OLS[a]		HLM[b]	
	Model 1	Model 2	Model 1	Model 2
Scaled score				
Urban	—	3.254 (2.704)	—	2.975 (2.512)
Size	—	–1.003 (1.085)	—	–0.900 (0.927)
% Minority	—	–3.131 (1.361)*	—	–3.168 (1.502)*
% Free lunch	—	–1.340 (1.367)	—	–1.345 (1.808)
Prior math	—	5.842 (1.166)*	—	5.859 (0.979)*
Treatment	6.945 (2.734)*	5.417 (2.083)*	6.750 (2.799)*	5.267 (2.030)*
Percent proficient				
Urban	—	0.004 (0.028)	—	0.004 (0.029)
Size	—	–0.004 (0.011)	—	–0.004 (0.010)
% Minority	—	–0.043 (0.014)*	—	–0.044 (0.014)*
% Free lunch	—	–0.023 (0.014)	—	–0.023 (0.015)
Prior math	—	0.071 (0.012)*	—	0.066 (0.012)*
Treatment	0.041 (0.032)	0.017 (0.022)	0.034 (0.033)	0.009 (0.020)

Note. OLS = ordinary least squares; HLM = hierarchical linear model; MEAP = Michigan Educational Assessment Program; Urban = metropolitan statistical area; Size= total school size; Minority = non-White student; Free lunch = free or reduced price lunch student; Prior math = average prior MEAP math achievement; ICC = intraclass correlation coefficient; HGLM = hierarchical generalized linear models. Model 1 only included a treatment variable. $ICC_{math\_scaled\_score} = 0.151$ and $ICC_{math\_proficiency} = 0.079$. Covariates are school level. Estimates shown here were run using HLM. HGLM and HLM results were comparable.

[a]Model includes 78 schools.

[b]Model includes 5,031 students in 78 schools.

*$p \leq .05$.

continuous measures of school achievement rather than proficiency metrics whenever possible.

Unfortunately, many states do not have this type of data available in their public-use data files. The appendix shows all 50 states and the metrics available. As of 2013, only 25 of the 50 states have publicly available scaled scores. The rest report only the percentage of students achieving various levels of proficiency, which means that in many states the confidence that could be placed in analyses based on the publicly available aggregate data would be limited. However, many states (43 of 50) do report the percentage of students reaching more than one cutoff score. For example, many states report the percentage of students achieving at a below basic, basic, proficient, and advanced level. These data could be used to look at the impact of a program for students at different points in the distribution, thereby reducing the likelihood of missing achievement gains that occur only at the tails of the distribution.

The lack of availability of scaled score metrics in state public-use data files is clearly one of the biggest limitations in using publicly available aggregate data in the evaluation of school-based interventions. Given the frequency with which these data could be used, the reduction in burden on both students and state personnel that this could achieve and the important information that could be learned about effective ways to intervene at the school level to improve student achievement by using such data, states should be encouraged to report not only the percent proficient metrics for their schools but also the average scaled scores.

Conclusion

Many researchers are reluctant to use aggregate school-level achievement data to evaluate

school-based interventions. As a result, considerable resources are often expended in an effort to obtain individual student-level data. However, we have shown that under a wide range of circumstances and for many key research questions, aggregate scaled scores are more than adequate for these purposes. First, aggregate school-level scaled scores appear to be quite robust for assessing the impact of an intervention in the context of a randomized experiment. HLM analyses with student-level data and OLS with school-level data provide unbiased estimates of program impact. Although in theory HLM protects against the influence of outliers, by providing precision weighted estimates, in practice, the results are very similar for HLM and OLS school-level models, under a wide range of possible scenarios, including large variations in the number of students per school. Simulations suggest that researchers only need be concerned when the number of schools in the sample is small (less than 30 schools) and the variation in the number of students per school is high. Even then, meaningful differences between the two estimation methods are found less than 1% of the time and are limited to samples in which the minimum enrollment is small (usually less than five). Our findings hold both in the presence of heterogeneous treatment effects and when school-level covariates are included.

Although one benefit of having access to individual student-level data is the availability of student-level covariates, individual student-level covariates do not appear to add much to the precision of the estimates that are obtained in our data. This finding is consistent with that of previous researchers (Bloom et al., 2007; Hedges & Hedberg, 2007; Konstantopoulos, 2012). And although one cannot, with data that are currently available, explore the heterogeneity of treatment effects based on students' prior achievement, because No Child Left Behind legislation requires that schools report disaggregate data based on a variety of other student characteristics, in some instances, aspects of student-level heterogeneity can be explored. Finally, we have argued that under many circumstances, aggregate school-level data are appropriate for nonexperimental analyses as well, although we were not able to explore these propositions empirically.

The biggest limitation to the use of aggregate data in the evaluation of school-based interventions has to do with the type of data that states are reporting. First, most states (30 out of 50 states in 2013) report only percent proficient metrics and not average scaled scores. As has been well documented, and as our simulations demonstrate, percent proficient metrics have many drawbacks that severely limit their utility in program evaluation. In particular, it is difficult to detect effects when the data are reported in terms of the percentage of students at or above proficiency and the intervention has effects on student achievement but does not move students across a state's proficiency threshold. Second, because of privacy concerns, states often restrict the reporting of aggregate scores to groups of a certain size. In some instances, these restrictions are quite large (as many as 40 students) and can severely limit the number of schools on which analyses can be conducted. Finally, states currently make only a limited number of potential covariates available at the school level, and this can have an impact on precision.

All states should be encouraged to make aggregate scaled scores publicly available. This would be a straightforward task and would save states and researchers a great deal of time and money. For researchers with Institutional Review Board (IRB) permission, states should also consider providing aggregate scaled scores for *all* schools and *all* subgroups regardless of size. This type of restricted-use data set would be easier to provide than student-level data and would provide researchers with a rich set of data with which to conduct analyses. Finally, states should be encouraged to make a wider range of aggregate variables available to researchers. In particular, including information disaggregated by students' level of prior achievement would considerably expand the range of analyses that researchers could conduct.

Appendix

Data Publicly Available as of February 2013

State[a]	Minimum number[b]	Scale	% proficient	Proficient level[c]	Grade	Subject[d]
Alabama	10	0	1	1	3–8, 11	Language, math, reading, science, social studies
Alaska	20, 40[e]	0	1	0	3–10	Math, reading, science, writing
Arizona	10	1	1	1	2–8, 10–12	Math, reading, science, writing
Arkansas	10	1	1	1	K–HS	ELA, math, science
California	10	1	1	1	2–11	ELA, history/social science, math, science
Colorado	16	1	1	1	3–10	Math, reading, science, writing
Connecticut	20	1	1	1	K, 3–8, 10	Math, reading, science, writing
Delaware	15	1	1	1	2–11	Math, reading, science, social science, writing
Florida	10	1	1	1	3–HS	Math, reading, science, social science, writing
Georgia	10	1	1	1	1–HS	ELA, math, reading, science, social studies, writing
Hawaii	40	0	1	0	3–HS	Math, reading, science, writing
Idaho	10	1	1	1	3–10	Language usage, math, reading, science
Illinois	10	0	1	1	3–8, 11	Math, reading, science
Indiana	10	1	1	0	3–HS	ELA, math, science, social studies
Iowa	10	0	1	1	3–8, 11	Math, reading, science
Kansas	10	0	1	1	3–8, 11–12	History, math, reading, science, writing
Kentucky	10	0	1	1	3–HS	Math, reading, science, social studies, writing
Louisiana	10	0	1	1	3–HS	ELA, math, science, social studies
Maine	20	1	1	1	3–8, 11	Math, reading, science, writing
Maryland	5	0	1	1	3–HS	Math, reading, science, government
Massachusetts	40	0	1	0	3–HS	ELA, math, science and technology/engineering
Michigan	30	1	1	1	3–9, 11	Math, reading, science, social studies, writing
Minnesota	10	1	1	1	3–HS	Math, reading, science
Mississippi	10	1	1	1	3–HS	ELA, math, science, social studies, writing
Missouri	30	1	1	1	3–HS	Communication arts, math, science, social studies
Montana	10	0	1	1	3–8, 10	Math, reading, science

(continued)

Appendix (continued)

State[a]	Minimum number[b]	Scale	% proficient	Proficient level[c]	Grade	Subject[d]
Nebraska	10	1	1	0	3–8, 11	Math, reading, science, writing
Nevada	10	0	1	1	3–8, 11	Math, reading, science, writing
New Hampshire	10	1	1	1	3–8, 11	Math, reading, science, writing
New Jersey	30	1	1	1	3–8, 11	LA, math, science
New Mexico	10	0	1	1	3–HS	Math, reading, science, social studies, writing
New York	5	1	1	1	3–HS	ELA, language, math, science, social studies, writing
North Carolina	5	1	1	1	3–HS	Math, reading, science, social science, writing
North Dakota	*ns*	0	1	1	3–8, 11	Math, reading, science
Ohio	10	0	1	1	3–8, 10–12	Math, reading, science, social studies, writing
Oklahoma	5	0	1	1	3–HS	Math, reading, science
Oregon	*ns*	0	1	1	3–8, 11	Math, reading, science, writing
Pennsylvania	40	0	1	1	3–8, 11	Math, reading, science, writing
Rhode Island	10	1	1	1	3–8, 11	Math, reading, science, writing
South Carolina	10	1	1	1	3–HS	ELA, math, science, social studies, writing
South Dakota	10	0	1	1	3–8, 11	Math, reading, science
Tennessee	10	0	1	1	3–HS	ELA, math
Texas	5	0	1	0	3–11	ELA, math, science, social studies, writing
Utah	10	0	1	0	3–HS	ELA, math, science
Vermont	*ns*	0	1	1	3–8, 11	Math, reading, science, writing
Virginia	10	1	1	1	3–HS	English, math, science, social studies, writing
Washington	30	0	1	1	3–HS	Math, reading, science, writing
West Virginia	10	0	1	1	3–11	Math, reading/LA, science, social studies
Wisconsin	6	1	1	1	3–8, 10	Math, reading, science, social studies
Wyoming	6	1	1	1	3–8, 11	Math, reading, science, writing
Total		25	50	43		

Note. K = kindergarten; HS = high school; ELA = English language arts, LA = language arts.

[a]Multiple links were utilized for several states.

[b]Minimum Number for Subgroup Reporting found on state applications, Principle 5.5. *ns* = states did not specify the minimum subgroup size in their state application.

[c]States providing proficiency-level data at multiple levels, for example, below standards, meets standards, and exceeds standards.

[d]All subjects are not available in all grade levels. Grade levels available may vary by year.

[e]Alaska 40 for limited English proficient and students with disabilities.

Authors' Note

Opinions expressed herein are those of the authors and do not represent the position of the U.S. Department of Education. The authors wish to thank Howard Bloom, Brian Jacob, and peer reviewers for helpful comments on the initial conceptualization and on earlier drafts, as well as Rachel Rifkin and Jessica Huff for their excellent research assistance. All errors and omission are our own.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was conducted as part of the School Leadership Improvement Study (SLIS). This research is supported by Grant R305A080696 from the Institute for Education Sciences (IES), U.S. Department of Education.

Notes

1. Later, we also consider the case of an ordinary least squares (OLS) estimate in which we weight by the number of students per school.

2. The average standard deviation of the control group in the simulated data is equal to 0.78 so that 0.05 of a standard deviation is equal to 0.039.

References

Bloom, H. S. (1984). Estimating the effect of job-training programs using longitudinal data: Ashenfelter's findings reconsidered. *Journal of Human Resources, 19*, 544–556.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*, 547−556.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*, 30–59.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724–750.

Gao, S., & Fraser, M. W. (2010). *Propensity score analysis*. Thousand Oaks, CA: SAGE.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science, 589*, 63–93.

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

Hill, J., Reiter, J., & Zanutto, E. (2004). A comparison of experimental and observational data analyses. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from an incomplete-data perspective* (pp. 49–60). New York, NY: John Wiley.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*, 351–360.

Holland, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3–17.

Kmenta, J. (1971). *Elements of econometrics*. New York, NY: Macmillan.

Konstantopoulos, S. (2012).The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research, 47*, 392–420.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129–149.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–264. http://dx.doi.org/10.2307/2533160

Somers, M. A., Zhu, P., Jacob, R., & Bloom, H. (2012, November). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation*. Paper presented at the American Public Policy and Management Association Fall Conference, Baltimore, MD.

Van den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement, 16*, 281–303.

65

Author Biographies

ROBIN T. JACOB is a research assistant professor at the University of Michigan's Institute for Social Research and the School of Education. She has a PhD in public policy from the University of Chicago. Her research focuses on evaluations of education interventions and evaluation methods. She has a special interest in how policies and program can affect instructional quality and outcomes in low-income elementary schools.

ROGER D. GODDARD is a Senior Fellow at McREL. His substantive interests include social cognitive theory, efficacy beliefs, and organizational leadership. His methodological expertise includes multilevel modeling, survey research, and experimental and quasi-experimental design.

EUN SOOK KIM, PhD, is an assistant professor in Measurement and Research in University of South Florida. Her research interest includes broadly quantitative data analysis and specifically, measurement invariance testing and multiple group analysis in structural equation modeling, multilevel modeling, and propensity score analysis.

Manuscript received March 1, 2012
Revision received December 6, 2012
Accepted January 31, 2013