Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Do Black Children Benefit More From Small Classes? Multivariate Instrumental Variable Estimators With Ignorable Missing Data

Yongyun Shin

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2012 37: 543
originally published online 27 April 2012
DOI: 10.3102/1076998611431083

The online version of this article can be found at:
http://jeb.sagepub.com/content/37/4/543

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
August 2012, Vol. 37, No. 4, pp. 543–574
DOI: 10.3102/1076998611431083

Do Black Children Benefit More From Small Classes? Multivariate Instrumental Variable Estimators With Ignorable Missing Data

Yongyun Shin
Virginia Commonwealth University

Does reduced class size cause higher academic achievement for both Black and other students in reading, mathematics, listening, and word recognition skills? Do Black students benefit more than other students from reduced class size? Does the magnitude of the minority advantages vary significantly across schools? This article addresses the causal questions via analysis of experimental data from Tennessee's Student/Teacher Achievement Ratio study where students and teachers are randomly assigned to small or regular class type. Causal inference is based on a three-level multivariate simultaneous equation model (SM) where the class type as an instrumental variable (IV) and class size as an endogenous regressor interact with a Black student indicator. The randomized IV causes class size to vary which, by hypothesis, influences academic achievement overall and moderates a disparity in academic achievement between Black and other students. Within each subpopulation characterized by the ethnicity, the effect of reduced class size on academic achievement is the average causal effect. The difference in the average causal effects between the race ethnic groups yields the causal disparity in academic achievement. The SM efficiently handles ignorable missing data with a general missing pattern and is estimated by maximum likelihood. This approach extends Rubin's causal model to a three-level SM with cross-level causal interaction effects, requiring intact schools and no interference between classrooms as a modified Stable Unit Treatment Value Assumption. The results show that, for Black students, reduced class size causes higher academic achievement in the four domains each year from kindergarten to third grade, while for other students, it improves the four outcomes except for first-grade listening in kindergarten and first grade only. Evidence shows that Black students benefit more than others from reduced class size in first-, second-, and third-grade academic achievement. This article does not find evidence that the causal minority disparities are heterogeneous across schools in any given year.

Keywords: *causal effect, reduced class size, ignorable missing data; instrumental variable, simultaneous equation model, Tennessee class size experiment*

1. Introduction

This article takes a new look at the implication of reduced class size for racial disparities in elementary school achievement analyzing the Tennessee's Student/Teacher Achievement Ratio study (STAR). The STAR randomly assigned teachers and students to small (13–17 classmates) or regular (22–25 classmates) class type and followed the students from kindergarten to third grade. By analyzing the experimental data, researchers have found strong evidence that reduced class size positively affects academic achievement (Finn & Achilles, 1990; Finn, Boyd-Zaharias, Fish, & Gerber, 2007; Krueger, 1999; Krueger & Whitmore, 2001; Mosteller, 1995; Nye, Hedges, & Konstantopoulos, 1999, 2000a; Shin & Raudenbush, 2011; Word et al., 1990;). Reducing class size, which aimed at improving academic achievement has been a popular educational policy of the federal government, states, and school districts for the last two decades (Milesi & Gamoran, 2006; Nye, Hedges, & Konstantopoulos, 2000b). Support for such a policy has heavily depended on the results from the class size studies of the STAR data (Finn & Achilles, 1990; Hanushek, 1999; Milesi & Gamoran, 2006). However, different investigators have drawn contradictory conclusions about whether Black students benefit more than other students from reduced class size. The author argues that recent advances in statistical analysis lay the basis for better answers to the questions than has been possible. A related aim of the article is to clarify these methodological advances in view of their potential application not only to this application but also to other data sets.

My primary substantive questions concern the impact on Black and other children of reduced class size in reading, mathematics, listening, and word recognition skills scores from Stanford Achievement Tests (SAT; Finn et al., 2007): What are the sizes and statistical significance of these effects? Are the effects for Black students significantly larger than the effects for others? Are the disparities heterogeneous across schools? Finn and Achilles (1990) analyzed the class mean achievement scores for White and minority students of the STAR and found that minority students benefit more from small class type than others in reading achievement, but not in mathematics and word recognition skills scores in first grade. Word et al. (1990) found no racial disparity in the effects of small class type on reading, mathematics, listening, and word recognition skills scores in their longitudinal analysis of the STAR data. Goldstein and Blatchford (1998) showed that, adjusting for kindergarten achievement scores, Black students exhibited larger effects of reduced class size on math and reading achievement scores than White counterparts in first grade. Krueger (1999) pooled the STAR data over the 4 years to analyze the effect of the intent-to-treat (ITT) random assignment to class types on the average percentile score of reading, math, and word recognition skills achievement separately for subsets of Black and White students. He found that Black students have a larger ITT effect than White students do. Nye et al. (2000b) reported that the positive effect of the ITT

assignment to a small class is not larger for minority students in either reading or mathematics achievement at any grade. Krueger and Whitmore (2001) compared the impacts of the ITT assignment to class types on the average percentile score of mathematics and reading tests for the full STAR sample and the subset of Black students. According to their analysis, the effects for the Black students are larger than those for all students each year from kindergarten to third grade. By likewise comparison, they also showed that attending a small class benefits Black students more than other students in terms of likelihood of taking a college entrance exam. Milesi and Gamoran (2006) analyzed the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K; Tourangeau, Nord, Lê, Sorongon, & Najarian, 2009) to find no differential benefit of attending a small class on reading and math achievement for kindergartners from different race ethnic backgrounds. These studies compared the effects of class types rather than the realized class size between student groups. Based on analysis of completely observed data, they require a strong assumption of data missing completely at random (MCAR; Rubin, 1976). The resulting estimators are inefficient and subject to bias (Little & Rubin, 2002).

To answer the causal questions, this article identifies the impact on academic achievement of class size using randomization to small and regular classes as an instrumental variable (IV; Angrist, Imbens, & Rubin, 1996). Class size and randomization operate at the classroom level while the outcome data vary at the student level. Two conventional approaches are two-stage least squares (2SLS) and maximum likelihood (ML). Krueger (1999) used the IV to show the significant effect of class size on the average of the percentile ranks of reading, math, and word recognition skills scores. Krueger and Whitmore (2001) used the same IV to find the significant effect of class size on the likelihood of taking a college entrance exam. Nye, Konstantopoulos, and Hedges (2004) also used the IV to identify the effect of class size on student achievement scores and gains. All these results are based on a univariate outcome analysis via 2SLS using completely observed cases or an ad hoc imputation of missing data such as a sample mean substitution. While useful, analysis of a single composite of exam scores may not reveal the impact of reduced class size on a subject-specific exam score while univariate analysis of exam scores one at a time treats the outcomes as if they are from different samples. Moreover, these conventional approaches are suboptimal in the presence of missing data, requiring a strong MCAR assumption. Shin and Raudenbush (2011) adapted ML to the case of three-level data with IVs, revealing explicitly the impacts of class size on multiple outcomes and allowing for efficient estimation under the comparatively weak assumption of ignorable missing data (Little & Rubin, 2002; Rubin, 1976). This article extends the ML approach to the analysis of multiple subpopulations characterized by student race ethnicity.

One of the intriguing claims based on STAR is that Black students benefit more than others from reduced class size, suggesting that reduced class size not

only increases academic achievement overall but also decreases racial inequality in academic achievement. However, the claim has been controversial (Finn & Achilles, 1990; Krueger, 1999; Krueger & Whitmore, 2001; Nye et al., 2000b; Word et al., 1990). This article aims to determine whether Black students benefit more than others from class size reduction. The causal analysis is based on a multilevel simultaneous equation model (SM) where the random class type is an IV, class size is an endogenous regressor, and both class type and class size interact with a Black student indicator. The IV causes class size to vary, which, by hypothesis, influences academic achievement overall and moderates a disparity in academic achievement between the two subpopulations of students. Within each subpopulation, the effects of reduced class size are the average causal effects and may be estimated by the method of Shin and Raudenbush (2011).

In the presence of ignorable missing data with a general missing pattern, the approach in this article analyzes all available data for efficient estimation of the racial inequalities in the causal effects of reduced class size on multiple outcomes. Methodological challenges arise in the causal analysis. First, efficient analysis requires that the causal impacts of reduced class size on multiple achievement scores for both Black and other students be simultaneously estimated. Analysis of one race ethnicity at a time not only produces inefficient estimators but also treats the ethnicity groups as if they are somewhat from different samples. Moreover, the assumption of data missing at random requires that all observed data be analyzed for efficient analysis (Little & Rubin, 2002; Rubin, 1976). Next, estimation of the minority disparity in the causal effects of class size amounts to estimation of the causal cross-level interaction effects between class size and Black student indicator on academic achievement when class size and class type operate at the classroom level. Third, the Rubin's Causal Model (RCM; Angrist et al., 1996; Holland, 1986; Rubin, 1978; Shin & Raudenbush, 2011) has to be extended to a multilevel SM where a quantitative mediator, class size, interacts with a Black student indicator. Furthermore, there is no widely available method for three-level missing data. This article extends the method of Shin and Raudenbush (2011) to handle ignorable missing data. Finally, the causal analysis is not complete without investigating if the causal disparities in multiple achievement outcomes between the two subpopulations of students vary randomly across schools that may be of different qualities.

Reduced class size may cause the minority disparities in academic achievement that are heterogenous across schools. Fryer and Levitt (2004) motivate this causal question. They analyzed the ECLS-K and found that Black students from kindergarten to first grade lose substantial ground in their test scores relative to White counterparts. If the gap were left to grow at the rate, by ninth grade, they predicted that the difference would be one full standard deviation in both math and reading test scores. They suggested that the relatively low quality of schools Black students attend may be to blame. A majority of the STAR schools were quite segregated, having their school percentages of Black students either 5%

or less, or greater than 95%. The author reasons that one of the main school quality indicators is how small the school's class sizes are. If reduced class size causes higher academic achievement for Black students than for others, which is heterogenous across schools of different qualities, it may provide important policy implication as poor school quality may be compensated by reduced class size. Such policymaking may lessen the ethnic gap in academic achievement.

Following Shin and Raudenbush (2011), this article refers to a simultaneous equation model as an SM to distinguish it from the popular mean and covariance structure method commonly denoted as a "structural equation model" or an "SEM" (Bollen, 1989). The next section clarifies the methodological approach for the causal inferences in a simple single-level context. The Data section describes the data for analysis. The Model section extends the single-level logic to multiple levels. Section 5 explains estimation with missing data. Section 6 shows the causal analysis. The Discussion section follows at last.

2. Single-Level Analysis

This section explains the multilevel causal analysis in a simple single-level SM where compliance to treatment assignment is perfect. The conventional SM approach makes it difficult to estimate the desired causal effects while the new approach in this article is straightforward. This section compares both approaches for the causal analysis. The new approach is then extended to the general case of multiple levels.

2.1. Conventional Approach

To show the difficulty involving the conventional SM method for the causal analysis, we may consider a simple desired structural SM of form

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 B_i + \beta_2 (S_i - \gamma_{s0}) + \beta_3 B_i (S_i - \gamma_{s0}) + u_i, \\ S_i &= \gamma_{s0} + \gamma_{s1} Z_i + a_{si}, \end{aligned} \tag{1}$$

where Y_i is a univariate exam score, B_i is a Black student indicator, class size S_i is an endogenous regressor, class type Z_i randomly assigned to students is an exogenous IV and $\begin{bmatrix} u_i \\ a_{si} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Delta_{yy} & \Delta_{ys} \\ \Delta_{ys} & \Delta_{ss} \end{bmatrix} \right)$ for student $i = 1, \ldots, n$. Class type is randomly assigned to a student at the beginning of the school year before school starts such that Z_i has the treatment effect γ_{s1} on the realized class size S_i without regard to race ethnicity. Exams are administered in the spring of the school year. The β_1 is the expected pretreatment gap in academic achievement between Black and other students. The desired causal parameters are β_2 and β_3 controlling for the pretreatment gap β_1. Reduced class size causes higher

academic achievement overall if both $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$ and benefits Black students more than others if $\beta_3 < 0$.

The reduced form of the structural SM (1) may be expressed as follows:

$$\begin{aligned} Y_i &= \gamma_{y0} + \gamma_{y1}B_i + \gamma_{y2}Z_i + \gamma_{y3}B_iZ_i + a_{y0i} + a_{y1i}B_i, \\ S_i &= \gamma_{s0} + \gamma_{s1}Z_i + a_{si}, \end{aligned} \tag{2}$$

where $\begin{bmatrix} a_{y0i} \\ a_{y1i} \\ a_{si} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \pi_{y0y0} & \pi_{y0y1} & \pi_{y0s} \\ \pi_{y0y1} & \pi_{y1y1} & \pi_{y1s} \\ \pi_{y0s} & \pi_{y1s} & \Delta_{ss} \end{bmatrix}\right)$. The structural SM (1) implies that $\gamma_{y0} = \beta_0$, $\gamma_{y1} = \beta_1$, $\gamma_{y2} = \beta_2\gamma_{s1}$, $\gamma_{y3} = \beta_3\gamma_{s1}$, $a_{y0i} = \beta_2 a_{si} + u_i$, and $a_{y1i} = \beta_3 a_{si}$ for $\pi_{y0y0} = \beta_2^2\Delta_{ss} + 2\beta_2\Delta_{ys} + \Delta_{yy}$, $\pi_{y0y1} = \beta_3(\beta_2\Delta_{ss} + \Delta_{ys})$, $\pi_{y0s} = \beta_2\Delta_{ss} + \Delta_{ys}$, $\pi_{y1y1} = \beta_3^2\Delta_{ss}$, and $\pi_{y1s} = \beta_3\Delta_{ss}$. Because $E(Y_i|Z_i = 1, B_i = 0) - E(Y_i|Z_i = 0, B_i = 0) = \gamma_{y2}$, $E(Y_i|Z_i = 1, B_i = 1) - E(Y_i|Z_i = 0, B_i = 1) = \gamma_{y2} + \gamma_{y3}$, and $E(S_i|Z_i = 1) - E(S_i|Z_i = 0) = \gamma_{s1}$ are causal effects induced by randomly assigned Z_i given B_i, so are the difference γ_{y3} in the causal effects and the IV estimands $\beta_2 = \gamma_{y2}/\gamma_{s1}$ and $\beta_3 = \gamma_{y3}/\gamma_{s1}$. We see that the unconstrained reduced-form SM (2) has 12 unique parameters while the structural SM (1) has only 9. The unconstrained model (2) identifies three extraneous parameters for subsequent analysis of the SM (1). To identify the desired SM (1), we have to impose constraints $\pi_{y1s} = \beta_3\Delta_{ss}$, $\pi_{y1y1} = \beta_3\pi_{y1s}$, and $\pi_{y0y1} = \beta_3\pi_{y0s}$ in Equation (2). Furthermore, Y_i and S_i may be subject to missingness. Consequently, it is difficult to estimate the SM (1) by the conventional 2SLS or ML.

When the pretreatment B_i takes b values for $b > 2$, structural SM (1) may be expressed as $Y_i = \sum_{j=0}^{b-1}[\beta_j + \beta_{b+j}(S_i - \gamma_{s0})]B_{ji} + u_i$ and the same S_i equation where $B_{0i} = 1$ and B_{ji} is an indicator of category $B_i = j$ for $j = 1, \ldots, b-1$. The interaction terms capture the class size-by-subgroup interaction effects. This approach becomes laborious with complicated constraints in estimation of the reduced-form SM.

2.2. New Approach

Because the randomized Z_i causes S_i to vary by $\gamma_{s0} + \gamma_{s1}Z_i$ on average, the $\gamma_{s1}Z_i$ is the causal effect of class type on class size centered around γ_{s0}. A desired structural SM for the causal analysis is also the SM (1) where $(\gamma_{s1}Z_i)$ replaces $(S_i - \gamma_{s0})$ in the first equation and everything else stays the same. The desired causal parameters are again β_2 and β_3 controlling for the pretreatment effect β_1. Then, the reduced-form equations are the SM (2) where $a_{y0i} + a_{y1i}B_i$ is replaced with a_{yi} and everything else stays the same. Therefore, simple transformations $\gamma_{y0} = \beta_0$, $\gamma_{y1} = \beta_1$, $\gamma_{y2} = \beta_2\gamma_{s1}$, $\gamma_{y3} = \beta_3\gamma_{s1}$ and $a_{yi} = u_i$ identify the desired structural SM (1).

Major advantages of this new approach over the conventional one are the simple transformation between structural and reduced-form SMs and the consequential straightforward estimation of the reduced-form SM without involving constraints. In addition, the advantages extend to applications where the covariate B_i has $b > 2$ values. Because the new SM approach is based on ML estimation of the multivariate reduced-form SM, it produces more efficient estimators than does 2SLS (Bollen, 1996; Imbens & Rubin, 1997a, 1997b; Little & Yau, 1998; Shin & Raudenbush, 2011). Furthermore, the new approach based on multivariate analysis facilitates use of all observed data for efficient analysis, given incomplete data under the comparatively weak assumption of ignorable missing data (Shin & Raudenbush, 2007, 2011) while 2SLS depends on complete-case analysis or an ad hoc imputation of missing data under the strong MCAR assumption.

To explain how the new approach handles missing data for efficient analysis when Y_i and S_i are subject to missingness, let I_n be an $n \times n$ identity matrix for a positive integer n and O_i denote an observed value indicator matrix for $[Y_i\ S_i]^T$ such that O_i is I_2 for both Y_i and S_i observed, $[1\ 0]$ for Y_i observed and S_i missing and $[0\ 1]$ for Y_i missing and S_i observed. In general, each row of O_i has a single one corresponding to the observed value and other elements equal to zero (Shin & Raudenbush, 2007, 2011). The O_i extracts all available data for efficient estimation. The ML estimation is based on the observed-data reduced-form SM

$$O_i \begin{bmatrix} Y_i \\ S_i \end{bmatrix} = O_i \begin{bmatrix} \gamma_{y0} + \gamma_{y1} B_i + \gamma_{y2} Z_i + \gamma_{y3} B_i Z_i + a_{yi} \\ \gamma_{s0} + \gamma_{s1} Z_i + a_{si} \end{bmatrix}. \tag{3}$$

3. Data

The Tennessee class size experiment was a study of one cohort of kindergartners in 1985 followed through third grade in 1989. Among all Tennessee schools invited to join, 79 kindergartens participated in the experiment with enough enrollment to create one of the three class types: small (13–17 students), regular (22–25 students), and regular-with-aide (22–25 students, a full-time teacher aide assigned to the class). In the fall of 1985 before school starts, 6,325 kindergartners along with their teachers were randomly assigned to 127 small, 99 regular, and 99 regular-with-aide classes. In the spring of 1986, they were assessed on reading, math, listening, and word recognition skills examinations from SAT. In the fall of 1986 before school starts, 2,314 new incoming first graders and 323 first-grade teachers were also randomly assigned to one of the three class types. In the spring of 1987, all first graders were assessed on the four exams. The new incoming 1,679 second and 1,283 third graders were randomized in the falls and assessed on the exams in the springs of the next two school years likewise. Students were to maintain their assigned class types until the end of third grade.

Some schools dropped out of the experiment leaving 76 schools for first grade and 75 schools for second and third grades. Despite the new incoming students each year, others left the participating schools leaving 6,829 first, 6,840 second, and 6,801 third graders. Consequently, some class sizes "drifted" from regular to small sizes and vice versa (Finn et al., 2007; Nye et al., 2000a). Out of a total of 11,601 students, only 26.6% remained in the experiment for the four consecutive years because of attrition and addition (Finn et al., 2007). Therefore, despite the intended small (13–17 students) and regular class sizes (22–25 students), the number of students for small and regular classes ranged 11–20 and 15–30, respectively. Although students stayed in the assigned class for their first year in the STAR experiment, some attended classes of different types from their ITT random assignments in subsequent years (Krueger, 1999; "Switching among class types," Nye et al., 2000a). For example, about 20% of kindergartners initially assigned to a regular or regular-with-aide class would attend a small class in a later year while 17% of those assigned to a small class would attend a class of the other types in subsequent years.

Although the randomization may sound questionable, studies have shown that it was successful (Finn et al., 2007; Krueger, 1999; Krueger & Whitmore, 2001; Nye, Konstantopoulos, & Hedges, 2004). Following the literature on class size (Finn & Achilles, 1990; Hanushek, 1999; Krueger, 1999; Krueger & Whitmore, 2001; Milesi & Gamoran, 2006; Nye et al., 1999, 2000a, 2000b), this article focuses on the analysis of small versus regular (including regular-with-aide) class types. For well-defined one IV per pupil, this study analyzes new students to STAR schools each year that include all 11,601 participants in the experiment (Shin & Raudenbush, 2011). This strategy satisfies some of the criteria used to negatively judge the validity of earlier results (Hanushek, 1999; Milesi & Gamoran, 2006; Nye et al., 1999). First, the approach removes a possible bias in the causal inference due to the "switching among class types" since the switching happened to some students in subsequent years following their first year at a STAR school. Also, the impact of each year's substantial attrition on the causal analysis is lessened (Goldstein & Blatchford, 1998; Krueger, 1999; Milesi & Gamoran, 2006). Furthermore, the approach eliminates potential sources of biased causal inference from differences in prior backgrounds between existing and new participants such as academic backgrounds and exposures to class types (Goldstein & Blatchford, 1998).

The Black student indicator is missing for 3, 29, 89, and 13 individuals (0%, 1%, 5%, and 1% of new students) who also miss 67%, 93%, 62%, and 52% of their achievement scores from kindergarten to third grade, respectively. Their within-class-type mean scores are different from the counterparts of all other new students in regular-class math and word recognition skills in second grade and small-class math in third grade. Among the second graders in regular classes, the students with missing race ethnicity have 0.63 and 0.43 standard deviations lower in the mean math and word recognition skills scores, respectively, than others.

Among the third graders in small classes, a single student misses the Black student indicator and has the math score higher than the mean score of other students. There exists no widely available three-level method that efficiently handles a mixture of continuous outcomes and a discrete covariate subject to missingness where the discrete covariate and another covariate interact to produce nonadditive effects on the outcomes. The causal analysis in this article drops these observations to maintain focus on the minority disparity in academic achievement induced by reduced class size. The overall consequence is to yield the conservative effects of reduced class size.

It is possible that the ad hoc deletion leads to the sample selection bias in the causal inferences. Later in this article, the causal analysis invokes the *no interference between classrooms* assumption such that the race ethnicity of a student may affect the potential outcomes of classmates, but of no others in any other classrooms or schools. Because 99%, 94%, 76%, and 97% of the classrooms from kindergarten to third grade, respectively, have completely observed race ethnicity, a great majority of the classrooms are not subject to the selection bias under this assumption. Furthermore, as the causal analysis shows later in this article, the random assignment of class type enables estimation of the treatment effects on potential class sizes without regard to the race ethnicity and on the potential academic achievement of a student controlling for no others' but her own race ethnicity. Therefore, the causal inferences in this article are arguably quite robust against the possible sample selection bias under the assumptions of the *no interference between classrooms* and the *random treatment assignment*. Notice that the deleted observations also result in loss of efficiency at the student level in the sense that they lead to reduced sample size only at the lowest level without affecting the number of classrooms and schools, and that key variables class size and class type operate at the classroom level. With many students for analysis each year, such loss practically does not seem to convey substantive importance in efficiency, in particular, for the causal effects of class size and class type that operate at the classroom level.

The data for analysis summarized in Table 1 consist of 6,322, 2,285, 1,590, and 1,270 new students attending 325, 322, 316, and 310 classes in 79, 76, 75, and 75 schools from kindergarten to third grade, respectively. Outcome variables are norm-referenced reading, math, listening, and word recognition skills scale scores from SAT that could be compared across years (Finn et al., 2007; Shin & Raudenbush, 2011). Small classes are 33% to 40%. Small and regular class sizes ranged 11–20 and 15–30 students, respectively. Black and female students are 33% to 46% and 46% to 49%, respectively. Out of non-Black students, 98% to 99% are White each year.

Missing achievement scores ranged 5% to 22%. One school in second grade missed all student achievement scores. One school in first grade and another in third grade had all reading scores missing. Seven, 23, and 25 classes in first, second and third grades, respectively, missed all achievement scores in at least one

TABLE 1
STAR Data for Analysis

Level	Variable	Description	Mean (Standard Deviation), missing %			
			Kindergarten	Grade 1	Grade 2	Grade 3
Student	B	1 if Black student	0.33 (0.47), 0%	0.38 (0.49), 0%	0.46 (0.50), 0%	0.40 (0.49), 0%
	READ	Reading score	437 (32), 8%	508 (51), 9%	572 (43), 22%	605 (37), 21%
	B=0		440 (32), 8%	520 (52),11%	587 (45), 26%	614 (37), 25%
	B=1		429 (29),10%	490 (42), 6%	556 (34), 17%	593 (33), 16%
	MATH	Math score	485 (48), 7%	522 (41), 5%	572 (41), 22%	608 (37), 20%
	B=0		491 (46), 7%	532 (40), 5%	586 (42), 26%	615 (37), 22%
	B=1		473 (49), 8%	508 (37), 4%	556 (34), 17%	597 (35), 17%
	LISTEN	Listening score	537 (33), 8%	562 (31), 6%	588 (34), 22%	617 (31), 21%
	B=0		545 (32), 7%	570 (31), 6%	603 (32), 26%	624 (30), 23%
	B=1		521 (30), 8%	548 (27), 4%	573 (28), 18%	607 (31), 18%
	WORD	Word recognition	434 (37), 7%	503 (51), 5%	571 (48), 11%	600 (43), 13%
	B=0	Skills score	438 (37), 7%	513 (50), 6%	584 (49), 13%	609 (43), 15%
	B=1		425 (34), 8%	488 (50), 4%	557 (43), 9%	587 (38), 11%
Class	S	Class size	19.5 (4.1), 0%	20.4 (4.0), 0%	20.4 (4.3), 0%	20.5 (4.5), 0%
	Z	1 if small class	0.39 (0.5), 0%	0.33 (0.5), 0%	0.36 (0.5), 0%	0.40 (0.5), 0%

of the four tests. Among these classes, 21 missed all student scores in all four outcomes. Complete-case analysis discards these schools and classes. Discarding a school, for example, amounts to dropping all its nested classes and students along with it from analysis. The analysis in this article uses these schools, classes, and students for efficient analysis because they have observed class size and class type and thus provide information at the school and classroom levels to strengthen inferences. All available achievement scores are analyzed to take advantage of their high correlations ranging from 0.5 to 0.91, which not only strengthen inferences in the presence of missing data but also make the assumption of ignorable missing data plausible (Shin & Raudenbush, 2011). Black students have less missing scores than others toward later years in the experiment. See Shin and Raudenbush (2011) for justification of the ignorable missing data assumption.

4. Model

This section extends the single-level logic in section 2 to multiple levels. Reduced class size, by hypothesis, causes higher academic achievement overall and moderates a disparity in academic achievement between Black and other students. The causal disparity is defined as the difference in the causal effects of reduced class size on academic achievement between the two subpopulations of students. First, the model assumptions are made explicit in the framework of RCM (Angrist & Imben, 1995; Angrist et al., 1996; Holland, 1986; Rubin, 1978). The single-level RCM is extended to a three-level SM where a quantitative mediator, class size, interacts with a Black student indicator. Next, the SM is described and then extended to a model where the causal minority disparity may randomly vary across schools.

4.1. Extended RCM

This section extends the RCM to a three-level SM having the continuous mediator whose value indicates the degree of compliance or the received treatment *dosage* and whose effect on academic achievement may differ across different race ethnic subpopulations of students. The RCM has been extended to a model involving more than two compliance statuses or treatments (Angrist & Imbens, 1995; Frangakis et al., 2004; Frangakis, Rubin, & Zhou, 2002; Imbens & Angrist, 1994). These approaches assumed monotonicity, which implies that students assigned to small class type have at least as small a class size as they would have had had they been assigned to regular class type. Small and regular class sizes of the STAR ranged 11 to 20 and 15 to 30, respectively, while their intended sizes were 13 to 17 and 22 to 25 students, respectively. Let $S_{jk}(1)$ and $S_{jk}(0)$ be potential class sizes for classroom j in school k, given small class type and regular class type, respectively, and let the classroom be assigned to regular class type. Although the classroom with class size $S_{jk}(0)$ may have had small class size

$S_{jk}(1)$ at least as small as the $S_{jk}(0)$ had it been assigned to small class type, it is hard to imagine such a mechanism that ensures $S_{jk}(1) \leq S_{jk}(0)$. On the other hand, likely mechanisms that may decrease a regular class size, for example, include parental complaints about large class sizes of their children, and reassignments of class types to correct behavioral problems, reward high-performing students, and compensate low-achieving students (Krueger, 1999). Such mechanisms may also increase small class sizes. Moreover, the peer impact of classmates switching class types may matter on class size that operates at the classroom level. In this scenario, a student assigned to small class type may experience her class size to drift larger due to the peer effect than the class size she would have had she been assigned to regular class type. Although the monotonicity assumption based on potential outcomes may not be tested, because we observe regular class sizes lower than some small class sizes and vice versa in sample and because plausible mechanisms exist that may make class sizes drift (Finn et al., 2007; Nye et al., 2000a), the monotonicity is not a plausible assumption for the causal analysis.

Shin and Raudenbush (2011) extended the RCM to a three-level SM with the continuous mediator, class size, and estimated the causal impact of reduced class size on academic achievement. To estimate the causal effect of taking algebra on eighth-grader's math achievement, Raudenbush (2010) took a 2SLS approach where assignment of a school to offering algebra is the IV, the indicator of taking or not taking algebra is the first mediator and peer ability is the second continuous mediator. He assumed "no compliance-effect covariance" to yield the unbiased estimator for the average causal effect of the peer ability on math achievement. The no compliance-effect covariance means no covariance between the effect of class-type assignment on class size and the effect of class size on academic achievement scores. Building on these advances, the modeling framework in this article extends the single-population analysis of Shin and Raudenbush to multiple subpopulations of students characterized by the race ethnicity.

To make the model assumptions explicit, let Z_{ijk} be 1 if class type is small and 0 otherwise where student i attends classroom j in school k for $i = 1, \ldots, n_{jk}$, $j = 1, \ldots, J_k$, and $k = 1, \ldots, K$. Then, $\mathbf{Z}_{jk} = (Z_{1jk}, \ldots, Z_{n_{jk}jk})$ are the class types of her n_{jk} classmates, $\mathbf{Z}_k = (\mathbf{Z}_{1k}, \ldots, \mathbf{Z}_{J_k k})$ of her $N_k = \sum_{j=1}^{J_k} n_{jk}$ schoolmates and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_K) = (Z_{ijk}, \mathbf{Z}_{-ijk})$ of $N = \sum_{k=1}^{K} N_k$ students in the entire sample where $\mathbf{Z}_{-ijk}$ denotes a vector of class types assigned of all other students except for her. Define $B_{ijk} = 1$ if she is an African American student and 0 otherwise and let A_{ijk} be the identification of the school to which she is assigned. Then, like class type, define $\mathbf{B}_{jk} = (B_{1jk}, \ldots, B_{n_{jk}jk})$ and $\mathbf{A}_{jk} = (A_{1jk}, \ldots, A_{n_{jk}jk})$ of her n_{jk} classmates, $\mathbf{B}_k = (\mathbf{B}_{1k}, \ldots, \mathbf{B}_{J_k k})$ and $\mathbf{A}_k = (\mathbf{A}_{1k}, \ldots, \mathbf{A}_{J_k k})$ of her N_k schoolmates and $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_K) = (B_{ijk}, \mathbf{B}_{-ijk})$ and $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_K)$ of all N students for the ethnicity and school assignment, respectively, so that

$S_{jk} = S_{jk}(\mathbf{Z}, \mathbf{B}, \mathbf{A})$ denotes her potential class size. Then, $\mathbf{S}_k = \mathbf{S}_k(\mathbf{Z}, \mathbf{B}, \mathbf{A}) = (S_{1k}, \ldots, S_{J_k k})$ are potential class sizes of J_k classes in her school and $\mathbf{S} = \mathbf{S}(\mathbf{Z}, \mathbf{B}, \mathbf{A}) = (\mathbf{S}_1, \ldots, \mathbf{S}_K)$ of all $J = \sum_{k=1}^{K} J_k$ classes such that $Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B}, \mathbf{A})$ are her potential achievement scores. This framework is very general that may involve all STAR students for analysis such that the class types and race ethnicities of other students may influence her potential outcomes. Causal inferences of class size with all STAR students may have to take this framework where ''switching among class types'' happens. For example, when a student switches class types, she carries her race ethnicity from her previous class type to the other. The consequence is that the new classroom that she attends gains not only one student in class size and her assigned class type but also one student in her race ethnicity. In such cases, the more classmates assigned to a small class or the lower her class proportion of the switching or both, the higher her academic achievement may be. Because the switching mechanism may involve $\mathbf{Z}_{-ijk}$ and $\mathbf{B}_{-ijk}$, an analyst has to consider their influence on the potential outcomes (Shin & Raudenbush, 2011; Verbitsky & Raudenbush, 2004).

Because the causal analysis involves newly entering STAR students each year who maintain their assigned class types, classmates share the same class type and schoolmates share the same school assignment so that $Z_{ijk} = Z_{jk}$ and $A_{ijk} = A_k$ for all i, for all j, and for all k. The estimands of interest are $E[S_{jk}(\mathbf{Z}, \mathbf{B}, \mathbf{A})] - E[S_{jk}(\mathbf{Z}', \mathbf{B}', \mathbf{A}')]$ and $E[Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B}, \mathbf{A})] - E[Y_{ijk}(\mathbf{S}', \mathbf{Z}', \mathbf{B}', \mathbf{A}')]$. The following assumptions extend the single-population approach of Shin and Raudenbush (2011) to the three-level causal inference involving multiple subpopulations (Angrist et al., 1996; Hong & Raudenbush, 2006; Imbens & Angrist, 1994; Shin & Raudenbush, 2011; Raudenbush, 2010):

1. *Intact schools:* It is unrealistic to think that students are randomly assigned to schools. Rather than dealing with unrealistic school assignments, the causal inferences in this article are to be generalized to a set of existing schools, focusing on the observed school assignments. Consequently, the potential outcomes of a student depend on the school that she is assigned to attend according to the observed school assignment. Thus, the causal estimands are estimated given observed school assignments $\mathbf{A} = \mathbf{a}$ in sample so that $S_{jk} = (\mathbf{Z}, \mathbf{B}, \mathbf{a} | \mathbf{A} = \mathbf{a}) = S_{jk}(\mathbf{Z}, \mathbf{B})$ and $Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B}, \mathbf{a} | \mathbf{A} = \mathbf{a}) = Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B})$;

2. *No interference between classes:* The potential class size for class j in school k does not depend on the treatment assignment and race ethnic composition of any other class in the same or other schools: $S_{jk}(\mathbf{Z}, \mathbf{B}) = S_{jk}(Z_{jk}, \mathbf{B}_{jk})$ for all $\mathbf{Z}$ and for all $\mathbf{B}$. The potential academic achievement for a student depends on the treatments assigned and received of, and race ethnic composition of classmates only, and of no others: $Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B}) = Y_{ijk}(S_{jk}, Z_{jk}, \mathbf{B}_{jk})$ for all $\mathbf{S}$, for all $\mathbf{Z}$, and for all $\mathbf{B}$;

3. *Exclusion restriction.* Class type affects potential academic achievement only through its effect on class size: $Y_{ijk}(\mathbf{S}, \mathbf{Z}, \mathbf{B}) = Y_{ijk}(\mathbf{S}, \mathbf{Z}', \mathbf{B})$ for all $\mathbf{S}$, for all $\mathbf{Z}$

and $\mathbf{Z}'$, and for all $\mathbf{B}$. By virtue of Assumptions 1 and 2, $Y_{ijk}(S_{jk}, Z_{jk}, \mathbf{B}_{jk}) = Y_{ijk}(S_{jk}, Z'_{jk}, \mathbf{B}_{jk})$ for all S_{jk}, for all Z_{jk} and Z'_{jk}, and for all $\mathbf{B}_{jk}$. That is, $Y_{ijk}(S_{jk}, Z_{jk}, \mathbf{B}_{jk}) = Y_{ijk}(S_{jk}, \mathbf{B}_{jk})$ for all S_{jk}, for all Z_{jk}, and for all $\mathbf{B}_{jk}$;

4. *Random treatment assignment.* Class type assignment Z_{jk} is random. As a result, $\mathbf{B}$ does not influence potential class sizes and, race-ethnic composition of other students does not impact her potential academic achievement on average. By virtue of Assumptions 1 and 2,

$$E[S_{jk}(1, \mathbf{B}_{jk}) - S_{jk}(0, \mathbf{B}_{jk})] = E[S_{jk}(1) - S_{jk}(0)],$$
$$E\{Y[S_{jk}(1), 1, \mathbf{B}_{jk}] - Y[S_{jk}(0), 0, \mathbf{B}_{jk}]\} = E\{Y[S_{jk}(1), 1, B_{ijk}] - Y[S_{jk}(0), 0, B_{ijk}]\}$$

for all $\mathbf{B}_{jk}$. Then, by the *Exclusion restriction,*

$$E\{Y[S_{jk}(1), 1, B_{ijk}] - Y[S_{jk}(0), 0, B_{ijk}]\} = E\{Y[S_{jk}(1), B_{ijk}] - Y[S_{jk}(0), B_{ijk}]\}$$

for all B_{ijk};

5. *Nonzero average causal effect of class type on class size.* The average causal effect of class type on class size is nonzero. By virtue of Assumptions 1, 2, and 4, $E[S_{jk}(Z_{jk}) - S_{jk}(Z'_{jk})] \neq 0$ for all $Z_{jk} \neq Z'_{jk}$;

6. *Linearity of academic achievement in class size*. By assumptions 1, 2 and 4, we may express potential outcomes as $S_{jk}(z)$ and $Y_{ijk}[S_{jk}(z), z, m]$ given $Z_{jk} = z$ and $B_{ijk} = m$ so that the ITT effects are

$$S_{jk}(1) - S_{jk}(0) = \Gamma_{s1jk}, \quad Y_{ijk}[S_{jk}(1), 1, m] - Y_{ijk}[S_{jk}(0), 0, m] = \Gamma_{y1ijkm}$$

for subscript m denoting $B_{ijk} = m$ given. Under this assumption of the linear dependence of Y_{ijk} on S_{jk},

$$Y_{ijk}[S_{jk}(1), m] - Y_{ijk}[S_{jk}(0), m] = \mathcal{B}_{1ijkm}[S_{jk}(1) - S_{jk}(0)] = \mathcal{B}_{1ijkm}\Gamma_{s1jk}.$$

Then, by the *exclusion restriction*, we have $Y_{ijk}[S_{jk}(z), z, m] = Y_{ijk}[S_{jk}(z), m]$ such that $\Gamma_{y1ijkm} = Y_{ijk}[S_{jk}(1), m] - Y_{ijk}[S_{jk}(0), m] = \mathcal{B}_{1ijkm}\ \Gamma_{s1jk}$. Let $E(\mathcal{B}_{1ijkm}) = \beta_{1m}$ and $E(\Gamma_{s1jk}) = \gamma_{s1}$. The sub-population average ITT effect given $B_{ijk} = m$ is

$$E(\Gamma_{y1ijkm}) = \gamma_{y1m} = \beta_{1m}\gamma_{s1} + cov(\mathcal{B}_{1ijkm}, \Gamma_{s1jk});$$

7. *No compliance-effect covariance.* This assumption says $\text{cov}(\mathcal{B}_{1ijkm}, \Gamma_{s1jk}) = 0$ to yield the unbiased IV estimand $\beta_{1m} = \gamma_{y1m}/\gamma_{s1}$ under Assumption 5. That is, there is no covariance between the impact of class type on class size and the impact of class size on achievement scores.

These assumptions extend the RCM to a three-level SM with multiple subpopulations characterized by B_{ijk}. The estimands of interest are $E[S_{jk}(1)] - E[S_{jk}(0)] = \gamma_{s1}$, the average *dosage*; $E[Y_{ijk}|Z = 1, B_{ijk} = m] - E[Y_{ijk}|Z = 0, B_{ijk} = m] = \gamma_{y1m}$, the subpopulation ITT effects; and $E\{Y_{ijk}[S_{jk}(1), m]\} - E\{Y_{ijk}[S_{jk}(0), m]\} = \beta_{1m}\gamma_{s1}$, the average subpopulation *dosage* effects on academic achievement. The IV estimand for the causal

effect of class size on academic achievement is $\beta_{1m} = \gamma_{y1m}/\gamma_{s1}$ (Shin & Raudenbush, 2011). Because class type was randomized within a school, it violates the *random treatment assignment* assumption and may have school-level confounders including school assignment. The causal analysis in this article assesses the influence of such confounding on the desired causal inferences.

4.2. Random-Intercepts SM

Reduced class size may cause higher academic achievement overall and moderate a disparity in academic achievement between Black and other students. A structural SM to address such causal inquiries is

$$\begin{aligned} Y_{ijk} &= \beta_0 + \beta_1 B_{ijk} + \beta_2(\gamma_{s1} Z_{jk}) + \beta_3 B_{ijk}(\gamma_{s1} Z_{jk}) + a_{yk} + b_{yjk} + \varepsilon_{ijk}, \\ S_{jk} &= \gamma_{s0} + \gamma_{s1} Z_{jk} + a_{sjk} + b_{sjk}, \end{aligned} \tag{4}$$

where Y_{ijk} is a vector of reading, math, listening, and word recognition skills test scores, B_{ijk} is a Black student indicator, class size S_{jk} is an endogenous regressor, class type Z_{jk} is randomly assigned class type to students, $(\gamma_{s1} Z_{jk})$ explains the causal variability in class size induced by Z_{jk}, $\begin{bmatrix} a_{yk} \\ a_{sk} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Delta_{yy} & \Delta_{ys} \\ \Delta_{sy} & \Delta_{ss} \end{bmatrix}\right)$, $\begin{bmatrix} b_{yjk} \\ b_{sjk} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Gamma_{yy} & \Gamma_{ys} \\ \Gamma_{sy} & \Gamma_{ss} \end{bmatrix}\right)$, and $\varepsilon_{ijk} \sim N(0, \boldsymbol{\Sigma})$ for student i attending classroom j in school k. Random effects are independent across different levels. The desired causal effects are β_2 and β_3 controlling for the pretreatment gaps β_1 in academic achievement. Reduced class size causes higher academic achievement overall if both $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$ and moderates a minority disparity of interest in academic achievement if $\beta_3 < 0$. As in the single-level case, Z_{jk} has the treatment effect γ_{s1} on the realized class size without regard to the race ethnicity by the *random treatment assignment*. However, compliance to treatment assignment that operates at the classroom level is not perfect. Because class type was randomized within each school violating the assumption, it may have school-level confounders. The causal analysis in this article assesses if causal inferences are biased due to such confounding later in this article. The IV Z_{jk} causes a nonzero effect on class size and affects academic achievement only through its effect on class size given B_{ijk}. That is, cov $(S_{jk}, Z_{jk}) \neq 0$ and $E(Z_{jk} a_{yk} | B_{ijk}) = E(Z_{jk} b_{yjk} | B_{ijk}) = E(Z_{jk} a_{sk}) = E(Z_{jk} b_{sjk}) = 0$.

To obtain the desired causal effects, let Equation (4) be reexpressed given $B_{ijk} = m$ as

$$\begin{aligned} Y_{ijk} &= \beta_{0m} + \beta_{1m}(\gamma_{s1} Z_{jk}) + a_{yk} + b_{yjk} + \varepsilon_{ijk}, \\ S_{jk} &= \gamma_{s0} + \gamma_{s1} Z_{jk} + a_{sk} + b_{sjk}, \end{aligned} \tag{5}$$

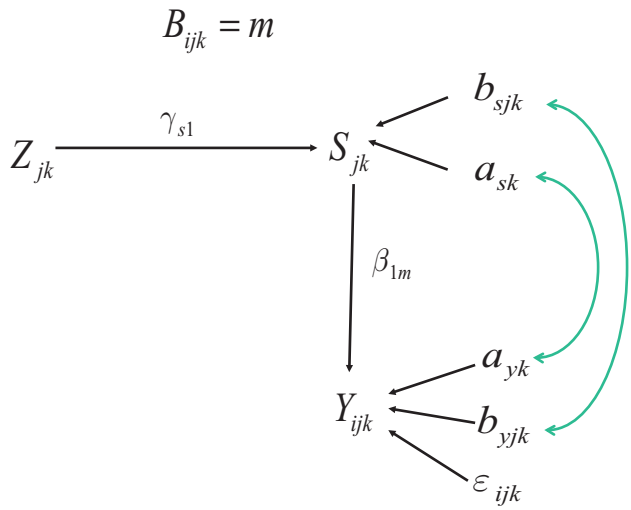

FIGURE 1. *Simultaneous Equation Model, SM (5) with an instrument variable (IV)* Z_{jk} *given* $B_{ijk} = m$.

for $\beta_0 = \beta_{00}, \beta_1 = \beta_{01} - \beta_{00}, \beta_2 = \beta_{10}, \beta_3 = \beta_{11} - \beta_{10}$ and $m = 0, 1$. The SM (5) shown in Figure 1 implies reduced-form equations

$$\begin{aligned} Y_{ijk} &= \gamma_{y0m} + \gamma_{y1m} Z_{jk} + a_{yk} + b_{yjk} + \varepsilon_{ijk}, \\ S_{jk} &= \gamma_{s0} + \gamma_{s1} Z_{jk} + a_{sk} + b_{sjk}, \end{aligned} \tag{6}$$

for $\gamma_{y0m} = \beta_{0m}$ and $\gamma_{y1m} = \beta_{1m}\gamma_{s1}$. The causal effects are $E[Y_{ijk}|Z_{jk} = 1, B_{ijk} = m] - E[Y_{ijk}|Z_{jk} = 0, B_{ijk} = m] = \gamma_{y1m}$ and $E[S_{jk}(1)] - E[S_{jk}(0)] = \gamma_{s1}$. SM (4) implies a reduced-form model

$$\begin{aligned} Y_{ijk} &= \gamma_{y0} + \gamma_{y1} B_{ijk} + \gamma_{y2} Z_{jk} + \gamma_{y3} B_{ijk} Z_{jk} + a_{yk} + b_{yjk} + \varepsilon_{ijk}, \\ S_{jk} &= \gamma_{s0} + \gamma_{s1} Z_{jk} + a_{sk} + b_{sjk}, \end{aligned} \tag{7}$$

for $\gamma_{y0} = \beta_0$, $\gamma_{y1} = \beta_1$, $\gamma_{y2} = \beta_2\gamma_{s1}$, and $\gamma_{y3} = \beta_3\gamma_{s1}$. SM (6) implies $\gamma_{y0} = \gamma_{y00}, \gamma_{y1} = \gamma_{y01} - \gamma_{y00}, \gamma_{y2} = \gamma_{y10}$ and $\gamma_{y3} = \gamma_{y11} - \gamma_{y10}$. The desired causal effects are $\beta_2 = \gamma_{y2}/\gamma_{s1}$ and $\beta_3 = \gamma_{y3}/\gamma_{s1}$.

4.3. Random-Coefficients SM

This section extends the random-intercepts SM (4) to an SM having random coefficients

$$
\begin{aligned}
Y_{ijk} &= (\beta_0 + u_{0k}) + (\beta_1 + u_{1k})B_{ijk} + (\beta_2 + u_{2k})(\gamma_{s1}Z_{jk}) + (\beta_3 + u_{3k})B_{ijk}(\gamma_{s1}Z_{jk}) + b_{yjk} + \varepsilon_{ijk}, \\
S_{jk} &= \gamma_{s0} + \gamma_{s1}Z_{jk} + a_{sk} + b_{sjk},
\end{aligned} \tag{8}
$$

where $\begin{bmatrix} u_{0k} \\ u_{1k} \\ u_{2k} \\ u_{3k} \\ a_{sk} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Delta_{00} & \Delta_{01} & \Delta_{02} & \Delta_{03} & \Delta_{0s} \\ \Delta_{10} & \Delta_{11} & \Delta_{12} & \Delta_{13} & \Delta_{1s} \\ \Delta_{20} & \Delta_{21} & \Delta_{22} & \Delta_{23} & \Delta_{2s} \\ \Delta_{30} & \Delta_{31} & \Delta_{32} & \Delta_{33} & \Delta_{3s} \\ \Delta_{s0} & \Delta_{s1} & \Delta_{s2} & \Delta_{s3} & \Delta_{ss} \end{bmatrix}\right)$ and others are defined the same as those in the SM (4). Random effects are again independent across levels. The SM implies that the IV Z_{jk} has a nonzero effect on class size and affects academic achievement only through its effect on class size given B_{ijk}. That is, $\text{cov}(S_{jk}, Z_{jk}) \neq 0$ and $E(Z_{jk}u_{lk}|B_{ijk}) = E(Z_{jk}b_{yjk}|B_{ijk}) = E(Z_{jk}a_{sk}) = E(Z_{jk}b_{sjk}) = 0$ for $l = 0, 1, 2, 3$. The reduced-form SM is

$$
\begin{aligned}
Y_{ijk} &= (\gamma_{y0} + a_{y0k}) + (\gamma_{y1} + a_{y1k})B_{ijk} + (\gamma_{y2} + a_{y2k})Z_{jk} + (\gamma_{y3} + a_{y3k})Z_{jk}B_{ijk} + b_{yjk} + \varepsilon_{ijk}, \\
S_{jk} &= \gamma_{s0} + \gamma_{s1}Z_{jk} + a_{sk} + b_{sjk},
\end{aligned} \tag{9}
$$

for $\gamma_{y0} = \beta_0$, $\gamma_{y1} = \beta_1$, $\gamma_{y2} = \beta_2\gamma_{s1}$, $\gamma_{y3} = \beta_3\gamma_{s1}$, $a_{y0k} = u_{0k}$, $a_{y1k} = u_{1k}$, $a_{y2k} = u_{2k}\gamma_{s1}$, and $a_{y3k} = u_{3k}\gamma_{s1}$. The effects induced by the randomly assigned Z_{jk} are $E(Y_{ijk}|Z_{jk} = 1, B_{ijk} = 0) - E(Y_{ijk}|Z_{jk} = 0, B_{ijk} = 0) = \gamma_2 + a_{y2k}$, $E(Y_{ijk}|Z_{jk} = 1, B_{ijk} = 1) - E(Y_{ijk}|Z_{jk} = 0, B_{ijk} = 1) = \gamma_{y2} + a_{y2k} + \gamma_{y3} + a_{y3k}$, their difference $\gamma_{y3} + a_{y3k}$ and $E(S_{jk}|Z_{jk} = 1) - E(S_{jk}|Z_{jk} = 0) = \gamma_{s1}$ given school-specific random effects a_{y2k} and a_{y3k}. The desired causal effects are $\beta_2 + u_{2k} = (\gamma_{y2} + a_{y2k})/\gamma_{s1}$ and $\beta_3 + u_{3k} = (\gamma_{y3} + a_{y3k})/\gamma_{s1}$ given a_{y2k} and a_{y3k}.

The SM (4) is nested within the SM (8) for $u_{1k} = u_{2k} = u_{3k} = 0$. The structural SM (8) yields 153 variance–covariance parameters at school level given less than 80 schools. To estimate the random minority disparities, this article analyzes the SM (8) having two outcomes at a time, (reading, math) followed by (listening, word recognitions skills).

5. Estimation With Missing Data

The reduced-form SMs (7) and (9) are estimated by the missing data method of Shin and Raudenbush (2011) that employs the expectation–maximization (EM) algorithm on variance components and Fisher scoring on fixed effects via ML (Dempster, Laird, & Rubin, 1977; Dempster, Rubin, & Tsutakawa, 1981; Laird & Ware, 1982; Longford, 1987). To sketch the method, let $a_{yk} = [a_{y0k}^T\ a_{y1k}^T\ a_{y2k}^T\ a_{y3k}^T]^T$ in the model (9), and let $a_k \sim N(0, \Delta)$, $b_{jk} \sim N(0, \Gamma)$, and $\varepsilon_{ijk} \sim N(0, \boldsymbol{\Sigma})$ for $a_k = [a_{yk}^T\ a_{sk}]^T$ and $b_{jk} = [b_{yjk}^T b_{sjk}]^T$ in the Equations (7) and (9). To relate $[Y_{ijk}^T\ S_{jk}]^T$ to the observed data, let O_{ijk} be the observed value indicator matrix for $[Y_{ijk}^T\ S_{jk}]^T$. Multiplication of the O_{ijk} to both sides of the

complete-data reduced-form SMs (7) and (9) yields the observed models. The O_{ijk} extracts all available data in sample for efficient analysis under the assumption of ignorable missing data. For the EM algorithm, $(Y_{ijk}, S_{jk}, a_k, b_{yjk})$ are viewed as complete data and $O_{ijk}[Y_{ijk}^T \; S_{jk}]^T$ observed for student i attending classroom j in school k in estimation of $(\gamma, \Delta, \Gamma, \mathbf{\Sigma})$ for $\gamma = [\gamma_{y0}^T \; \gamma_{y1}^T \; \gamma_{y2}^T \gamma_{y3}^T \; \gamma_{s0} \; \gamma_{s1}]^T$ (Shin & Raudenbush, 2011).

The next section illustrates how to make the desired causal inferences. The desired SMs are estimated by the author's C program via ML. The convergence criterion is the difference in the observed log-likelihoods between two consecutive iterations taken to be less than 10^{-6}. The statistical significance of an effect estimate is discussed at a significance level .05.

6. Analysis

This section illustrates the causal analysis beginning with the causal ITT effect on academic achievement. The model is the Y_{ijk} equation of the reduced-form SM (7) and is called a "3L ITT," a three-level ITT model to assess the causal impact of the ITT intervention to treat a student to reduced class size controlling for the pretreatment effect of race ethnicity. Next, the structural SM (4) is estimated to study if reduced class size causes higher academic achievement overall and moderates a disparity in academic achievement between Black and other students. This model is referred to as a "3L Random Int.," a three-level random-intercepts model. The analysis then extends to estimation of three-level random-coefficients SM (8), "3L Random Coef." These models are named after the comparable models of Shin and Raudenbush (2011) which facilitates the comparison. Because class type was randomly assigned within a school, it violates the *random treatment assignment* assumption and may have school-level confounders. To assess if such confounding seriously biases the causal inferences, an alternative model is estimated and compared that controls for all school-level covariates, both observed and unobserved, by fixed school effects (Shin & Raudenbush, 2011).

Model diagnostics[1] identified an outlying mathematics score 288 of a female African American kindergartner in a small class who missed all other exams. The outlier, when included in analysis, lowered the causal effect estimates on all outcomes up to 5%. This section presents the analysis of 6,321 kindergartners without the outlier. The minimum mathematics score in kindergarten is now 320.

6.1. ITT Causal Effects

This analysis examines if the ITT intervention to treat a student to reduced class size causes higher academic achievement overall and moderates a disparity in academic achievement between Black and other students. The ITT model is the first equation in the SM (7) where the desired causal effects are $E[Y_{ijk}|Z_{jk} =$

$1, B_{ijk} = 0] - E[Y_{ijk}|Z_{jk} = 0, B_{ijk} = 0] = \gamma_{y2}$ and $E[Y_{ijk}|Z_{jk} = 1, B_{ijk} = 1] - E[Y_{ijk}|Z_{jk} = 0, B_{ijk} = 1] = \gamma_{y2} + \gamma_{y3}$. Their difference γ_{y3} is the causal minority disparity in academic achievement induced by the randomized ITT intervention. The results under "3L ITT" in Table 2 may be compared to the causal effects of Z_{jk} in Shin and Raudenbush (2011) under "SR 3L ITT" which is the 3L ITT for $\gamma_{y1} = \gamma_{y3} = 0$. The SR 3L ITT shows that the ITT treatment causes higher academic achievement in all test subjects throughout the 4 years except for second-grade math, listening, and word recognition skills.

The γ_{y1} of the 3L ITT displays significant pretreatment minority gaps in all test subjects throughout the 4 years (Finn & Achilles, 1990; Fryer & Levitt, 2004; Goldstein & Blatchford, 1998; Krueger, 1999; Word et al., 1990). For non-Black students, the ITT treatment causes higher academic achievement in all subjects but first-grade listening in kindergarten and first grade only, while for Black students, it causes higher academic achievement in all subjects throughout the 4 years, controlling for the effects of pretreatment race ethnicity. The minority disparities are pronounced in second and third grades and on first-grade listening achievement. The ITT effects for Black students corresponding to the significant disparities are 3 to 19 times as large as their counterparts for other students in magnitude. The likelihood ratio tests for $\gamma_{y3} = 0$ yield p values .82, .04, .24, and .09 from kindergarten to third grade, respectively, in modest to strong support of the 3L ITT in first to third grade.

The 3L ITT is subject-specific. Given race ethnicity, a Black third grader assigned to reduced class size, for example, improves his or her math achievement score by 17.89 points on average. The improvement is comparable to the corresponding expected pretreatment minority gap in math achievement, 19.6 points lower than that of a non-Black student. The subject-specific results may have policy implications that lead to interventions targeting specific test subjects.

To visually compare the 3L ITT effects across race ethnicity groups in Table 2, Figure 2 draws the expected achievement scores against class types given a test subject, a grade, and race ethnicity where K, G1, G2, and G3 stand for kindergarten and first to third grades, respectively. The legend in the first graph applies to all graphs and indicates that a solid line connects the expected scores in small and regular classes for a non-Black student while a dotted line links those for a Black student. A 95% confidence interval for each expected score is also drawn vertically. In kindergarten and first grade, both Black and other students have better academic achievement in small classes than they do in regular classes throughout all subjects on average although the G1 listening improvement for non-Black students looks relatively weak. In second and third grades, Black students continue to exhibit better expected academic achievement for every subject in small classes than they do in regular classes while such improvement seems to disappear in every subject for non-Black students. Consequently, the pretreatment racial gap in academic achievement evident throughout all 4 years in Table 2

TABLE 2
Analysis of ITT Effects.

		3L ITT				SR 3L ITT	2L ITT			
Outcome	Yr.	γ_{y1}	γ_{y2}	γ_{y3}	$\gamma_{y2} + \gamma_{y3}$	γ_{y2}	γ_{y1}	γ_{y2}	γ_{y3}	$\gamma_{y2} + \gamma_{y3}$
READ	K	−13(1)*	4.97(1.12)*	1.08(1.69)	6.05(1.53)*	5.31(0.98)*	−13(2)*	4.92(1.40)*	1.59(2.22)	6.52(1.96)*
	1	−21(2)*	7.81(2.73)*	0.07(4.18)	**7.87(3.45)***	7.82(2.27)*	−15(3)*	7.36(3.60)*	−0.67(5.63)	**6.70(4.52)**
	2	−23(2)*	2.40(2.79)	**6.68(4.02)***	9.09(3.11)*	4.04(2.13)*	−14(4)*	2.13(3.72)	**6.55(5.31)**	8.69(3.91)*
	3	−23(2)*	−0.82(2.20)	12.11(3.61)*	11.30(2.94)*	5.25(1.84)*	−19(4)*	1.73(3.07)	9.39(5.03)*	11.12(4.06)*
MATH	K	−22(2)*	8.47(1.77)*	−1.01(2.59)	7.46(2.38)*	8.14(1.56)*	−23(2)*	8.39(2.22)*	−0.14(3.43)	8.25(3.06)*
	1	−20(2)*	8.83(2.18)*	−2.10(3.31)	6.73(2.74)*	7.93(1.82)*	−20(3)*	9.43(2.83)*	−3.27(4.42)	6.16(3.56)*
	2	−27(2)*	−3.14(2.80)	11.29(4.01)*	8.15(3.09)*	0.91(2.20)	−24(4)*	−2.55(3.74)	13.11(5.27)*	10.57(3.85)*
	3	−20(2)*	2.55(2.40)	15.33(3.90)*	17.89(3.25)*	9.00(2.03)*	−19(4)*	4.48(3.21)	14.65(5.27)*	19.14(4.32)*
LISTEN	K	−24(1)*	2.89(1.12)*	1.41(1.73)	4.30(1.54)*	3.40(0.99)*	−23(2)*	2.54(1.41)*	2.40(2.28)	4.93(1.99)*
	1	−21(1)*	2.28(1.68)	7.43(2.57)*	9.71(2.12)*	5.00(1.43)*	−19(2)*	1.67(2.22)	8.59(3.45)*	10.26(2.78)*
	2	−22(2)*	1.85(2.23)	2.34(3.18)	**4.18(2.47)***	2.09(1.70)	−15(3)*	1.56(2.96)	2.96(4.18)	**4.52(3.08)**
	3	−18(2)*	0.31(2.09)	**7.97(3.39)***	8.28(2.84)*	4.22(1.77)*	−18(3)*	2.02(2.74)	**7.17(4.54)**	9.18(3.73)*
WORD	K	−14(1)*	5.56(1.27)*	1.01(1.95)	6.57(1.75)*	5.88(1.11)*	−13(2)*	5.58(1.59)*	1.40(2.57)	6.98(2.24)*
	1	−17(2)*	9.79(2.86)*	−1.12(4.34)	**8.68(3.60)***	9.31(2.36)*	−13(3)*	9.23(3.73)*	−2.11(5.82)	**7.12(4.69)**
	2	−21(3)*	0.50(3.15)	9.71(4.56)*	10.21(3.53)*	3.42(2.37)	−12(5)*	−0.52(4.20)	10.45(6.03)*	9.93(4.47)*
	3	−22(2)*	0.00(2.51)	12.48(4.12)*	12.48(3.36)*	6.20(2.09)*	−15(5)*	2.82(3.47)	11.15(5.69)*	13.97(4.60)*

Note: Significant effects are marked by asterisks. Estimates (standard errors) resulting in different statistical inferences between 3L ITT and 2L ITT are boldfaced.

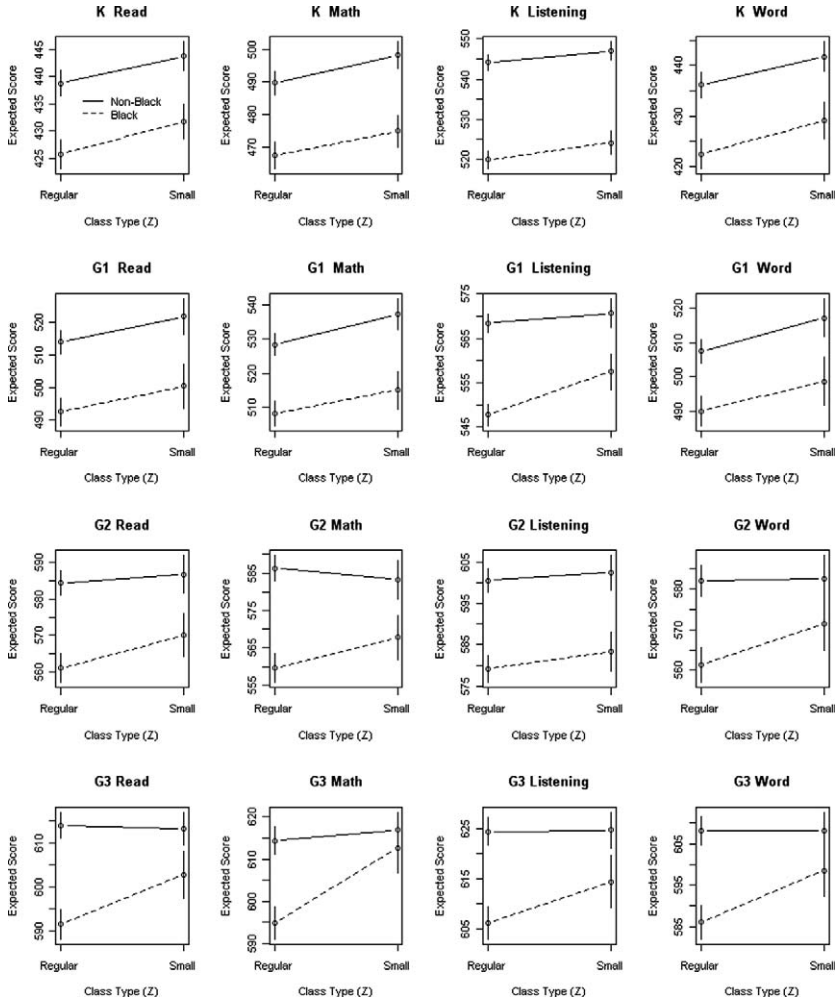

FIGURE 2. *Each graph given a test subject and a grade draws expected scores against class types. The legend in the first graph applies to all graphs and shows that the solid and dotted lines connect the expected scores in regular and small classes of non-Black and Black students, respectively. A vertical line represents a 95% confidence interval for each expected score.*

seems to considerably reduce in small classes. For example, the pretreatment racial gap of 19.6 points in G3 math achievement seems to decrease so substantially in small classes that one half of the 95% confidence interval (607, 619) for the expected math score of a Black student overlaps the counterpart (613, 621) of a non-Black student.

To rule out a potential problem of biased inferences due to confounding between Z_{jk} and school-level covariates, the 3L ITT may be compared to an alternative model

$$Y_{ij} = \gamma_{y0} + \gamma_{y1}B_{ij} + \gamma_{y2}Z_j + \gamma_{y3}B_{ij}Z_j + \gamma_{y4}A_j + b_{yj} + \varepsilon_{ij}, \quad (10)$$

where A_j is a vector of school indicators having fixed effects γ_{y4}, $b_{yj} \sim N(0, \Gamma_{yy})$, $\varepsilon_{ij} \sim N(0, \boldsymbol{\Sigma})$ and all others are defined in the same way as the three-level counterparts in the SM (7) for student $i = 1, \ldots, n_j$ attending classroom $j = 1, \ldots, J$. The two-level model (10), free of school-level confounders, is called a 2L ITT, a two-level ITT model. If the confounders seriously bias the causal inferences, the causal estimates between 2L ITT and 3L ITT will be different. The resulting estimates are shown under 2L ITT in Table 2 where those with conflicting statistical inferences between the two models are boldfaced. The relative inefficiency of the 2L ITT is notable by the larger standard error for every estimate, up to 100% larger. The pretreatment minority gaps under γ_{y1} are all significant with the magnitudes comparable to or smaller than the three-level counterparts. The differences in magnitude are due to confounding with school-level covariates. The 2L ITT effects γ_{y2} for non-Black students are comparable to their counterparts resulting in the same statistical inferences. The 2L ITT disparities γ_{y3} are likewise comparable to their three-level counterparts except for the statistically insignificant 6.55 (5.31) on second-grade reading and 7.17 (4.54) on third-grade listening compared to the significant three-level counterparts 6.68 (4.02) and 7.97 (3.39). The 2L ITT effects $\gamma_{y2} + \gamma_{y3}$ for Black students are comparable to their counterparts except for the statistically insignificant 6.70(4.52) and 7.12(4.69) on first-grade reading and word recognition skills and 4.52(3.08) on second-grade listening compared to the significant three-level counterparts 7.87(3.45), 8.68(3.60), and 4.18(2.47), respectively. With comparable estimates between the two models, these conflicting statistical inferences are mainly due to the relative inefficiency of the two-level analysis. Overall, the confounding between class type and school-level confounders does not seriously bias the causal inferences.

6.2. Causal Effects of Reduced Class Size

This analysis examines if reduced class size causes higher academic achievement overall and moderates a disparity in academic achievement between Black and other students. The desired model is the SM (4) where the desired causal effects are β_2 and $\beta_2 + \beta_3$ for non-Black and Black students controlling for the pretreatment minority gap β_1 in academic achievement. Their differences β_3 are the causal disparities induced by reduced class size. The results are displayed under 3L Random Int. in Table 3. For comparison, the results under "SR 3L Random Int." show the analysis of the SM (4) for $\beta_1 = \beta_3 = 0$ in Shin and Raudenbush (2011).

TABLE 3

3L Random Int., SM (4); SR 3L Random Int., SM (4) with $\beta_1 = \beta_3 = 0$; and 2L Fixed, SM (11).

		3L Random Int.				SR 3L Random Int.	2L Fixed			
Outcome	Yr.	β_1	β_2	β_3	$\beta_2 + \beta_3$	β_2	β_1	β_2	β_3	$\beta_2 + \beta_3$
READ	K	−13(1)*	−0.68(0.15)*	−0.14(0.23)	−0.82(0.21)*	−0.72(0.13)*	−13(2)*	−0.67(0.19)*	−0.21(0.30)	−0.88(0.27)*
	1	−22(2)*	−1.09(0.38)*	−0.02(0.58)	**−1.11(0.48)***	−1.09(0.32)*	−15(3)*	−1.03(0.50)*	0.09(0.79)	**−0.94(0.63)**
	2	−23(2)*	−0.32(0.35)	−0.79(0.50)	−1.11(0.39)*	−0.52(0.27)*	−14(4)*	−0.29(0.47)	−0.76(0.67)	−1.05(0.49)*
	3	−23(2)*	0.08(0.27)	−1.46(0.44)*	−1.38(0.36)*	−0.66(0.23)*	−18(4)*	−0.23(0.38)	−1.15(0.62)*	−1.38(0.50)*
MATH	K	−22(2)*	−1.16(0.24)*	0.15(0.35)	−1.01(0.32)*	−1.11(0.21)*	−23(2)*	−1.15(0.30)*	0.03(0.47)	−1.11(0.42)*
	1	−20(2)*	−1.24(0.31)*	0.35(0.46)	**−0.88(0.38)***	−1.09(0.26)*	−20(3)*	−1.33(0.40)*	0.52(0.62)	**−0.80(0.50)**
	2	−26(2)*	0.46(0.35)	−1.53(0.50)*	−1.07(0.39)*	−0.11(0.27)	−24(4)*	0.32(0.47)	−1.65(0.66)*	−1.33(0.49)*
	3	−19(2)*	−0.32(0.30)	−1.97(0.48)*	−2.29(0.40)*	−1.14(0.25)*	−19(4)*	−0.58(0.40)	−1.82(0.65)*	−2.40(0.54)*
LISTEN	K	−24(1)*	−0.40(0.15)*	−0.18(0.24)	−0.58(0.21)*	−0.46(0.14)*	−23(2)*	−0.35(0.19)*	−0.32(0.31)	−0.66(0.27)*
	1	−21(1)*	−0.31(0.23)	−1.06(0.36)*	−1.38(0.30)*	−0.70(0.20)*	−19(2)*	−0.23(0.31)	−1.23(0.48)*	−1.46(0.39)*
	2	−21(2)*	−0.22(0.28)	−0.28(0.40)	−0.50(0.31)	−0.25(0.21)	−15(3)*	−0.22(0.37)	−0.30(0.52)	−0.52(0.39)
	3	−18(2)*	−0.04(0.26)	**−1.02(0.42)***	−1.07(0.35)*	−0.54(0.22)*	−18(3)*	−0.27(0.34)	**−0.88(0.56)**	−1.15(0.46)*
WORD	K	−14(1)*	−0.76(0.17)*	−0.13(0.27)	−0.89(0.24)*	−0.80(0.15)*	−13(2)*	−0.76(0.22)*	−0.18(0.35)	−0.95(0.31)*
	1	−17(2)*	−1.36(0.40)*	0.14(0.60)	**−1.22(0.50)***	−1.29(0.33)*	−13(3)*	−1.28(0.52)*	0.28(0.81)	**−1.01(0.65)**
	2	−20(3)*	−0.10(0.39)	−1.15(0.57)*	−1.25(0.44)*	−0.46(0.29)	−11(5)*	0.05(0.53)	−1.28(0.76)*	−1.23(0.56)*
	3	−22(2)*	−0.02(0.31)	−1.51(0.51)*	−1.53(0.41)*	−0.78(0.26)*	−15(5)*	−0.36(0.43)	−1.36(0.70)*	−1.72(0.57)*

Note: Significant effects are marked by asterisks. Estimates (standard errors) resulting in different statistical inferences between 3L Random Int. and 2L Fixed are boldfaced.

The SR model shows that reduced class size causes higher academic achievement in all test subjects throughout the 4 years except for second-grade math, listening, and word recognition skills.

The pretreatment minority gaps in academic achievement under the β_1 of the 3L Random Int. are all statistically significant with the estimates practically identical to the counterparts of the 3L ITT. For non-Black students, reduced class size causes higher academic achievement in all test subjects except for first-grade listening in kindergarten and first grade only, while for Black students, it causes higher academic achievement in all test subjects throughout all years with a relatively modest effect on second-grade listening, controlling for the pretreatment effects of race ethnicity. In second and third grades when the causal disparities under β_3 are most pronounced, the effects of class size for Black students are 2 to 76 times as large as the counterparts for other students in magnitude. The likelihood ratio tests for $\beta_3 = 0$ produce p values .83, .03, .18, and .08 from kindergarten to third grade, respectively. The hypothesis tests are not elaborate on significant effects. The Bonferroni multiple comparisons for $\beta_3 < 0$ at a family-wise significance level .1 reveal that Black students benefit more from reduced class size than other students in terms of mathematics and word recognition skills scores in second grade as the 3L Random Int. displays. Therefore, the hypothesis tests support that reduced class size benefits Black students more than other students in terms of academic achievement from first to third grade.

The results are subject-specific. Given student race ethnicity, reducing class size by the sample average *dosage* of 8.03 classmates in third grade, for example, causes a Black student to improve her math achievement by 18.39 (2.29×8.03) points on average. The magnitude is similar to the corresponding ITT effect, 17.89 points, of the 3L ITT. Other effects are similarly comparable between 3L Random Int. and 3L ITT. Although the *exclusion restriction* assumption based on potential outcomes is not testable, the similarities in the comparison strengthen its plausibility (Shin & Raudenbush, 2011). The estimate 18.39 is also comparable to her pretreatment minority gap, 19.6 points lower than the math score of a non-Black student on average.

Figure 3 reveals the differential impacts of class size on academic achievement between Black (dotted line) and other students (solid line). Given a test subject and a grade, each graph draws expected achievement scores against class size centered around the mean regular class size. The class size ranges from 12 to 28 supported by sample class sizes each year. The mean regular class sizes are 22.2, 22.9, 23.3, and 23.6 from kindergarten to third grade, respectively. As class size decreases in kindergarten and first grade, academic achievement improves on average for both race ethnicity groups in all subjects although such improvement looks relatively weak for first-grade listening achievement of non-Black students. In second and third grades, as class size reduces, Black students continue to perform better in all subjects while non-Black students exhibit relatively trivial or no improvement in all subjects on average.

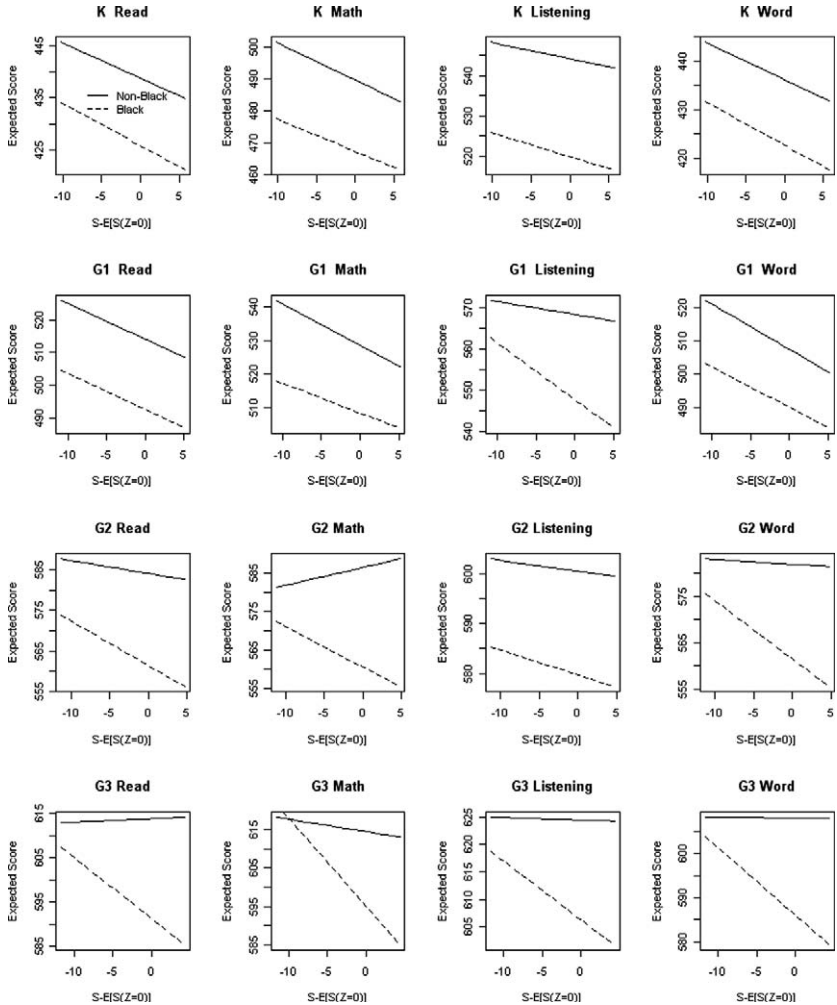

FIGURE 3. *Graphs are drawn in the same way as those of Figure 2 except for class size centered around mean regular class size on the horizontal axis.*

To rule out a potential problem of biased causal inferences due to confounding between Z_{jk} and school-level covariates, the 3L Random Int. is compared to an alternative model controlling for school effects

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 B_{ij} + \beta_2(\gamma_{s1} Z_j) + \beta_3 B_{ij}(\gamma_{s1} Z_j) + \beta_4 A_j + b_{yj} + \varepsilon_{ij}, \\ S_j &= \gamma_{s0} + \gamma_{s1} Z_j + \gamma_{s2} A_j + b_{sj}, \end{aligned} \tag{11}$$

where A_j is a vector of school indicators having fixed effects β_4 and γ_{s2} on Y_{ij} and S_j, respectively, $\begin{bmatrix} b_{yj} \\ b_{sj} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Gamma_{yy} & \Gamma_{ys} \\ \Gamma_{sy} & \Gamma_{ss} \end{bmatrix}\right)$, $\varepsilon_{ij} \sim N(0, \boldsymbol{\Sigma})$ and all others are defined in the same way as the three-level counterparts in the SM (4) for student $i = 1, \ldots, n_j$ attending classroom $j = 1, \ldots, J$. The equation is called 2L Fixed, a two-level model with school fixed effects. The estimates are displayed under 2L Fixed in Table 3 where those with conflicting statistical inferences between 3L Random Int. and 2L Fixed are boldfaced. The pretreatment minority gaps under β_1 are comparable to or lower than the counterparts of 3L Random Int. in magnitude. The differences are due to school-level confounders. For example, while the school percentage of Black students is greater than 95% for 16 schools, it is 5% or less for more than 30 schools each year. Thus, a majority of the STAR schools were quite segregated. The β_2 estimates produce the same statistical inferences between 3L Random Int. and 2L Fixed. The statistical inferences for β_3 are identical under the two models except for the insignificant $-0.88(0.56)$ on third-grade listening achievement under 2L FIXED compared to the significant three-level counterpart $-1.02(0.42)$. The $(\beta_2 + \beta_3)$ estimates for Black students result in the same statistical inferences under the two models except for the insignificant $-0.94(0.63)$, $-0.80(0.50)$, and $-1.01(0.65)$ on first-grade reading, math and word recognition skills under 2L FIXED compared to the significant three-level counterparts $-1.11(0.48)$, $-0.88(0.38)$, and $-1.22(0.50)$. With comparable estimates, the conflicting statistical inferences are mainly due to the relative inefficiency of the two-level approach. Overall, no serious sign of bias is indicated in the causal inferences based on the 3L Random Int.

6.3. Heterogenous Disparities

The 3L Random Int. in Table 3 reveals that the disparities in the causal effects of reduced class size are pronounced in favor of Black students in second and third grades and on first-grade listening achievement. The causal disparities may randomly vary across schools if Black students attend schools of low qualities relative to the schools that other students attend (Fryer & Levitt, 2004). The different school qualities are plausible as a majority of the STAR schools are quite segregated. The random-coefficients SMs (8) are estimated two outcomes at a time, (reading, math) and (listening, word recognition skills).

The estimated models appear in Table 4. The pretreatment minority gaps are as strong as those of the 3L Random Int. Overall, the β_2 and β_3 estimates are comparable to the 3L Random Int. counterparts. From the variance estimates of Δ_{33}, the minority disparities seem modestly heterogenous across schools in kindergarten and first grade. The likelihood ratios testing $u_{1k} = u_{2k} = u_{3k} = 0$ for (reading, math) and (listening, word recognition skills) outcome pairs produce p values .03 and .40 for kindergarten, .84 and .70 for first grade, .70 and .66 for second grade, and 1.00 and 1.00 for third grade. The likelihood ratio

TABLE 4
Three-Level Random Coefficients SM (8)

Outcome	Gr.	β_1	β_2	β_3	Diagonal Elements of Δ_{33}
READ	K	−13.2(1.6)	−0.75(0.16)	−0.08(0.27)	2.48(1.81)
	1	−23.6(2.5)	−0.99(0.42)	0.30(0.66)	10.87(9.67)
	2	−26.9(2.7)	0.19(0.42)	−1.08(0.61)	10.26(9.68)
	3	−23.8(2.2)	0.38(0.32)	−2.36(0.50)	1.08(6.05)
MATH	K	−21.7(2.4)	−1.16(0.25)	0.26(0.42)	5.41(4.08)
	1	−20.5(2.1)	−1.10(0.34)	0.66(0.54)	9.75(6.61)
	2	−29.1(2.6)	0.78(0.43)	−1.48(0.60)	11.15(9.18)
	3	−20.4(2.4)	−0.10(0.33)	−3.02(0.53)	1.80(6.44)
LISTEN	K	−23.7(1.5)	−0.40(0.17)	−0.19(0.26)	1.40(1.69)
	1	−20.0(1.6)	−0.13(0.25)	−0.85(0.41)	5.32(3.83)
	2	−21.5(2.4)	−0.40(0.32)	0.29(0.43)	1.58(4.65)
	3	−18.1(2.1)	0.01(0.25)	−1.61(0.44)	1.44(4.51)
WORD	K	−12.5(1.7)	−0.76(0.19)	−0.02(0.29)	1.41(2.04)
	1	−17.0(2.6)	−1.16(0.41)	0.69(0.68)	9.71(9.76)
	2	−22.0(3.0)	−0.18(0.44)	−0.36(0.65)	11.19(11.15)
	3	−22.5(2.5)	−0.09(0.34)	−1.33(0.52)	0.65(6.93)

testing $u_{3k} = 0$ for the (reading, math) outcome pair in kindergarten yields a p value equal to .23. Therefore, the analysis does not find evidence that reduced class size induces the minority disparities that are heterogenous across schools.

With 75 to 79 schools each year, however, the school-level random effects produce 45 variance covariance components so that weak power to detect the random effects is consequential. The resulting uncertainty in estimation of many parameters may have contributed to imprecise estimation of the random effects.

7. Discussion

The analysis in this article extended the Rubin's causal modeling framework to a three-level SM having a continuous mediator whose value indicates the degree of compliance or the received treatment *dosage* and whose effects on the outcome variables may differ across multiple subpopulations of students. The extension enabled this study to find that for Black students, reduced class size causes higher academic achievement in reading, math, listening, and word recognition skills throughout the 4 years from kindergarten to third grade, while for non-Black students, reduced class size causes higher academic achievement in the four outcomes except for first-grade listening in kindergarten and first grade only. Hypothesis tests revealed that Black students benefit more from reduced class size than others in terms of academic achievement in first, second, and third grades. The analysis was then extended to a three-level random-coefficients SM where the minority

disparities in the causal effects of reduced class size on academic achievement were hypothesized to be heterogenous across schools. This article did not find evidence that the minority disparities varied randomly across schools.

The causal analysis in this article is based on seven assumptions. Cases may be made to violate each assumption (Shin & Raudenbush, 2011). For the *no compliance-effect covariance* assumption, for example, if teachers who are used to teaching small classes are more likely to teach small classes better, they bias the causal impact of reduced class size on academic achievement. If students with prior exposure to a certain class type are more likely to learn better in the class type, they bias the causal effect, too. The assumptions, however, seem reasonable within the context of the current application. The *intact schools* assumption is realistic with existing school assignments. The *no interference between classes* assumption seems reasonable because students share academic experience with classmates most. The *random treatment assignment* assumption was violated due to the randomization within schools. This violation was shown to yield no serious bias in the causal inferences. The *exclusion restriction* assumption is reasonable because randomly labeling each student by class type cannot affect academic achievement unless it induces the *dosage* in class size. The *nonzero average causal effect of class type on class size* assumption is very reasonable from the sample average *dosage* greater than 7 each year. The *no compliance-effect covariance* assumption seems plausible from the fact that both students and teachers were randomly assigned to class type so that their differences in ability to learn and teach are also randomized across class types. Consequently, the violating cases of this assumption above are unlikely. School differences due to randomization within schools have been shown to cause no serious bias in the causal inferences.

The *no compliance-effect covariance* assumption based on potential outcomes is not directly testable. However, a testable implication of the assumption exists by the principal stratification (Frangakis & Rubin, 2002). One may expect the estimated IV estimands, the principal effects, to differ across the principal strata if the assumption is violated. As a simple example, two principal strata may consist of two sets of classes with potential outcomes $S_{jk}(1) > S_{jk}(0)$ and $S_{jk}(1) < S_{jk}(0)$ where classes with $S_{jk}(1) = S_{jk}(0)$ do not need to be considered by the *exclusion restriction*. The principal strata are not affected by treatment and hence considered as a pretreatment covariate (Frangakis & Rubin, 2002). If we denote $S_+ = \{j : S_{jk}(1) < S_{jk}(0)\}$ and $S_- = \{j : S_{jk}(1) > S_{jk}(0)\}$, the comparison between $\{Y_{ijk}(1) : S_s\}$ and $\{Y_{ijk}(0) : S_s\}$ produces a causal effect within the stratum $s =$ '+' or '−'. Frangakis and Rubin (2002) showed how to predict the missing membership of an individual to a principal stratum and the missing potential outcomes $Y(z)$ for a single-level analysis that may be extended to three-level data. Such an analysis is beyond the scope of the current article.

Missing achievement scores were handled according to the efficient missing data method of Shin and Raudenbush (2011). This method can efficiently handle

ignorable missing data with a general missing pattern at any of the levels under the normal theory. However, sample data may have nonignorable missing patterns (Rubin, 1976; Little & Rubin, 2002). For example, low performing students may be more likely to miss exams than high performing counterparts. The ignorable missing data assumption is then violated. A sensitivity analysis to the assumption may be a valuable future research topic.

Acknowledgments

The author appreciates helpful comments from the reviewer. The author also thanks Dr. Stephen W. Raudenbush for making helpful comments on an earlier version of this manuscript and Dr. Jeremy D. Finn for providing the STAR data.

Declaration of Conflicting Interests

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

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D090022 to NORC. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

Note

1. Based on the posterior distributions of random effects given observed data, the residuals at all levels of each SM were obtained after convergence. An influential mathematics score 288 on the fitted models of a female African American kindergartner in a small class was identified at level 1. The analysis with the outlier lowered statistically significant causal effect estimates on all outcomes up to 5% compared to one without. Consequently, the student was dropped from analysis, and this article presents the analysis of 6,321 kindergartners without the outlier. Assumed normality and linearity looked reasonable at all levels.

References

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the Acoustical Society of America*, *90*, 431–442.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the Acoustical Society of America*, *91*, 444–455.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K. A. (1996). An alternative two stage least squares estimator for latent variable equations. *Psychometrika*, *61*, 109–121.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, *76*, 1–38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the Acoustical Society of America*, *76*, 341–353.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Ecucational Research Journal*, *27*, 557–577.

Finn, J. D., Boyd-Zaharias, J., Fish, R. M., & Gerber, S. B. (2007). *Project STAR and beyond: Database user's guide*. Lebanon, TN: HEROS.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Mahboobeh, S., Valhov, D., & Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the Acoustical Society of America*, *99*, 239–249.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics, 58,* 21–29.

Frangakis, C. E., Rubin, D. B., & Zhou, X. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, *3*, 147–164.

Fryer, R. G. Jr., & Levitt, S. D. (2004). Understanding the Black-White test score gap in the first two years of school. *Review of Economics and Statistics*, *86*, 447–464.

Goldstein, H., & Blatchford, P. (1998). Class size and eduational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal*, *24*, 255–268.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee's STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21,* 143–163.

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

Hong, G., & Raudenbush, S. W. (2006). Evaluating Kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the Acoustical Society of America*, *101*, 901–910.

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

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for cuasal effects in randomized experiments with noncompliance. *Annals of Statistics*, *25*, 305–327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, *64*, 555–574.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, *114*, 497–532.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project STAR. *Economic Journal*, *111*, 1–28.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963–974.

Lintz, M., Folger, J., & Breda, C. (1990). The state of Tennessee's Student/Teacher Achievement Ratio (STAR) project: Final summary report 1985–1990. Nashville: Tennessee State Department of Education.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York, NY: Wiley.

Little, R. J. A., & Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods, 3*, 147–159.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74,* 817–827.

Milesi, C., & Gamoran, A. (2006). Effects of class size and instruction on kindergarten achievement. *Educational Evaluation and Policy Analysis*, *28*, 287–313.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children: Critical Issues for Children and Youths*, *5*, 113–127.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis*, *21*, 127–142.

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

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000b). Do the disadvantaged benefit more from small classes? Evidence from the Tennessee class size experiment. *American Journal of Education*, *109*, 1–26.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257.

Raudenbush, S. W. (2010). *Strategies for modeling interference between units in multisite trials*. New Orleans, LA: Presentation at ENAR.

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Shin, Y., & Raudenbush, S. W. (2007). Just-identified versus over-identified two-level hierarchical linear models with missing data. *Biometrics*, *63*, 1262–1268.

Shin, Y., & Raudenbush, S. W. (2011). The causal effect of class size on academic performance: Multivariate instrumental variable estimators with Tennessee class size data missing at random. *Journal of Educational and Behavioral Statistics*, *36*, 154–185.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, Kindergarten class of 1998-99 (ECLS-K), Combined user's manual for the ECLS-K Eighth-Grade and K-8 Full sample data files and electronic codebooks (NCES 2009-004)*. Washington, DC: NCES, IES, DOE.

Verbitsky, N., & Raudenbush, S. W. (2004). Causal inference in spatial setting. *Preceedings of the Social Statistics Section, American Statistical Association, Social Statistics Section [CD-ROM],* Alexandria, VA: American Statistical Association, 2369–2374.

Word, E., Johnston, J., Bain, H., Fulton, B., Zaharias, J., Achilles, C., Lintz, M., Folger, J., & Breda, C. (1990). The state of Tennessee's student/teacher achievement ratio (STAR) project: Final summary report 1985–1990. Nashville: Tennessee State Department of Education.

Author

YONGYUN SHIN is an assistant professor, Department of Biostatistics at Virginia Commonwealth University, 830 East Main Street, P.O. Box 980032, Richmond, VA 23298-0032; email: yshin@vcu.edu. His research areas are efficient analysis of hierarchical models given incomplete data, statistical computing and design of experiment.

Manuscript received May 7, 2010
Revision received February 8, 2011
Accepted May 20, 2011