Article

Evaluation Review
2016, Vol. 40(4) 334-377

DOI: 10.1177/0193841X16671283
journals.sagepub.com/home/erx

Meaningful Effect Sizes, Intraclass Correlations, and Proportions of Variance Explained by Covariates for Planning Two- and Three-Level Cluster Randomized Trials of Social and Behavioral Outcomes

Nianbo Dong[1], Wendy M. Reinke[1], Keith C. Herman[1], Catherine P. Bradshaw[2], and Desiree W. Murray[3]

Abstract

Background: There is a need for greater guidance regarding design parameters and empirical benchmarks for social and behavioral outcomes to inform assumptions in the design and interpretation of cluster randomized

[1] University of Missouri, Columbia, MO, USA
[2] University of Virginia, Charlottesville, VA, USA
[3] University of North Carolina, Chapel Hill, NC, USA

Corresponding Author:
Nianbo Dong, University of Missouri, 14 Hill Hall, Columbia, MO 65203, USA.
Email: dong.nianbo@gmail.com

trials (CRTs). **Objectives:** We calculated the empirical reference values on critical research design parameters associated with statistical power for children's social and behavioral outcomes, including effect sizes, intraclass correlations (ICCs), and proportions of variance explained by a covariate at different levels (R^2). **Subjects:** Children from kindergarten to Grade 5 in the samples from four large CRTs evaluating the effectiveness of two classroom- and two school-level preventive interventions. **Measures:** Teacher ratings of students' social and behavioral outcomes using the Teacher Observation of Classroom Adaptation–Checklist and the Social Competence Scale–Teacher. **Research design:** Two types of effect size benchmarks were calculated: (1) normative expectations for change and (2) policy-relevant demographic performance gaps. The ICCs and R^2 were calculated using two-level hierarchical linear modeling (HLM), where students are nested within schools, and three-level HLM, where students were nested within classrooms, and classrooms were nested within schools. **Results and Conclusions:** Comprehensive tables of benchmarks and ICC values are provided to inform prevention researchers in interpreting the effect size of interventions and conduct power analyses for designing CRTs of children's social and behavioral outcomes. The discussion also provides a demonstration for how to use the parameter reference values provided in this article to calculate the sample size for two- and three-level CRTs designs.

Keywords
cluster randomized trials (CRTs), empirical benchmarks, effect size, design parameters, social and behavioral outcomes, statistical power

Cluster randomized trials (CRTs) are now widely used to examine intervention effects in prevention science (National Research Council & Institute of Medicine, 2009). In order to design CRTs with sufficient statistical power to detect a meaningful effect size of an intervention with precision, researchers need to make reasonable assumptions about the design parameters to estimate the required sample sizes. In addition to the discretionary factors that are based on the researcher's judgment (e.g., type I error, one- or two-tailed test), several inherent factors that depend on the nature of the intervention and the study design, and are outside of the researcher's control are associated with statistical power in CRTs: (1) effect sizes, (2) intraclass correlations (ICCs), (3) proportions of variance explained by a covariate at different levels (R^2), and (4) sample size

(Bloom, Richburg-Hayes, & Black, 2007; Hedges & Rhoads, 2010; Konstantopoulos, 2008; Raudenbush, 1997; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008). The larger the desired/expected effect size, the greater the statistical power. However, the assumption of the desired effect size of an intervention needs to be realistic, which could be small but still meaningful. Statistical power is sensitive to ICC, and it decreases when ICC increases. Yet including the covariates that are correlated with the outcome can improve estimate precision and increase statistical power (Bloom et al., 2007; Raudenbush et al., 2007). Statistical power increases when the sample size increases. When a researcher conducts power analysis of a CRT to estimate the sample size that is needed to detect the intervention effect with sufficient power (e.g., 80%), three critical design parameters (effect sizes, ICCs, and R^2) need to be carefully chosen and justified.

Recently, several studies have reported the ICCs and R^2 for academic achievement outcome measures (e.g., Hedges & Hedberg, 2007, 2013, on mathematics and reading; Westine, Spybrook, & Taylor, 2013, on science achievement) and outcome measures for teacher professional development (Kelcey & Phelps, 2013). However, there is limited information about these design parameters on social and behavioral outcomes to inform the design of CRTs.

It is helpful to use empirical benchmarks for interpreting effect size in prevention science. The effect size (i.e., the standardized mean difference, calculated by the difference of the means between the treatment and control groups, divided by the pooled standard deviation of the two groups; we used Hedges' g as the effect size metric for the interventions in this article) provides one way to interpret the substantive or practical significance as compared to the statistical significance of interventions (Bloom, Hill, Black, & Lipsey, 2008; Hill, Bloom, Black, & Lipsey, 2008; Lipsey et al., 2012). Small effects are expected from universal preventive interventions, given that they are delivered to the entire populations without regard to risk (i.e., many individuals would not develop disorders even without the intervention); yet even very small effects on a population level can result in dramatic improvements in public health outcomes (National Research Council & Institute of Medicine, 2009). However, Bloom, Hill, Black, and Lipsey (2008), Hill, Bloom, Black, and Lipsey (2008), and Lipsey et al. (2012) similarly argued that effect sizes should be interpreted with respect to empirical benchmarks that are relevant to the intervention, target population, and outcome measure being considered. In particular, these benchmarks can include (a) normative expectations for change, (b) policy-relevant performance gaps, and (c) effect size results from similar

studies. Bloom, Hill, and colleagues (2008) illustrated these benchmarks regarding academic achievement. But to date, there are no comparable studies of social and behavioral measures that would provide empirical benchmarks necessary for researchers and policy makers to interpret the magnitude of the prevention effects on social and behavioral outcomes.

Social and behavioral measures are commonly used in social science research as primary and secondary outcomes of interest. In particular, teacher ratings of student social skills and emotional and behavior problems provide an important perspective about youth development. Despite recent criticisms of the use of questionnaires to gauge educational outcomes (Duckworth and Yeager, 2015), teacher ratings of youth emotional and behavior problems have proven remarkably predictive of long-term student outcomes. For instance, teacher ratings of attention problems and disruptive behaviors in Grade 1 predict likelihood of school dropout (Johns Hopkins Prevention Intervention Research Center (JHU PIRC), 2006), violence (Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004), and criminality (Schaeffer et al., 2006) in adolescence. Moreover, teacher ratings are sensitive to intervention effects over time, even with different teacher informants (see Bradshaw, Mitchell, & Leaf, 2010; Kellam et al., 2011). Although using multiple informants and methods to measure any construct is preferred given the limitations of any single method or source (Duckworth & Yeager, 2015), teacher ratings provide a unique and important component of any school-based research measurement scheme.

The purpose of the current study was 2-fold: (1) to provide empirical benchmarks for researchers and policy makers to interpret the magnitude of the intervention effects on social and behavioral outcomes and (2) to provide reference values of these design parameters on social and behavioral outcomes for researchers to conduct power analysis of CRTs. Our sample was drawn from four large-scale cluster random assignment designs of school-based prevention programs targeting social and behavioral outcomes in children. Our three-level data, whereby students (Level 1) were nested within classrooms (Level 2), and classrooms were nested within schools (Level 3), allow us to conduct analyses using three-level hierarchical linear modeling (HLM) as well as two-level HLM where students are nested within schools by ignoring classroom level. These data were used to estimate (1) the empirical benchmark of meaningful effect sizes regarding the developmental annual change, and policy-relevant demographic performance gaps (race/ethnicity, gender, and socioeconomic status [SES]); (2) ICCs for schools and classrooms; and (3) R^2 of demographic and pretest covariates at different levels (school, classrooms, and student). We conclude with recommendations for how the current findings could be used for

planning CRTs of similarly designed social and behavioral outcomes for school-aged children. These findings contribute to the literature by providing empirical benchmarks for interpreting the magnitude of the intervention effects on social and behavioral outcomes and reference design parameters values to inform power analyses for two- and three-level CRTs.

Method

Sample

Data for this study came from four separate large projects that used cluster random assignment designs to evaluate the effectiveness of two different school-based prevention interventions. Two projects evaluated the efficacy of the Incredible Years Teacher Classroom Management (IY TCM) program, which is a universal classroom management training program for teachers (Murray, Rabiner, & Carrig, 2014; Reinke, Herman, & Dong, 2014). The other two projects tested a school-wide prevention model called Positive Behavioral Interventions and Supports (PBIS), which aims to improve student behavior through improved school climate and whole-school behavior management practices (Bradshaw, Koth, Thornton, & Leaf, 2009; Bradshaw, Pas, Goldweber, Rosenberg, & Leaf, 2012). All four projects included a primary outcome of teacher reports of students' behavior using a checklist style teacher rating system. The projects, samples, and measure are described in greater detail below.

Project 1 tested IY TCM and included 105 teachers and 1,818 students in kindergarten to Grade 3 from nine urban Missouri schools serving primarily African American students (Reinke et al., 2014). Teachers within schools were randomly assigned to receive IY TCM or to a wait-list control group. Data for the present analyses were collected at the fall of the school year (baseline, prior to the intervention), and the late spring of the school year (posttest) on two outcome measures: the Teacher Observation of Classroom Adaptation–Checklist (TOCA-C; Koth, Bradshaw, & Leaf, 2009; Werthamer-Larsson, Kellam, & Wheeler, 1991) and the Social Competence Scale–Teacher Scale (T-COMP; Conduct Problems Prevention Research Group, 2002).

Project 2 also tested IY TCM (Murray et al., 2014) and enrolled 1,276 students in kindergarten to Grades 2 from 97 classrooms across 11 schools. Grade levels within schools were randomly assigned to receive the IY TCM program or to a wait-list comparison condition. Participating schools were drawn from three rural to semirural school districts in the central part of North Carolina. The data for Project 2 analyses were collected in the fall of

the school year (baseline, prior to the intervention) and late spring of the school year (posttest) on the T-COMP (Conduct Problems Prevention Research Group, 2002).

Project 3 focused on school-wide PBIS using a randomized effectiveness trial design to test the universal elements of the intervention relative to a control. Specifically, the CRT randomly assigned 37 Maryland elementary schools in five school districts to either a treatment or control condition (Bradshaw et al., 2009, 2010). Participating schools were diverse with regard to setting and economic characteristics; 48% were suburban, 41% urban fringe, and 49% received Title I support. The trial included 2,596 school staff members (1,437 general education teachers and 1,159 support staff including school counselors and psychologists) and 12,341 students. These data were collected at the fall of the school year (baseline) following the initial summer training intervention and late spring of the first school year (Posttest 1) and three follow-up years (Posttests 2, 3, and 4) on the TOCA-C (Koth et al., 2009; Werthamer-Larsson et al., 1991).

Project 4 involved a CRT, whereby all schools were implementing the universal Tier 1 elements of PBIS, but approximately half of the schools were randomly assigned to implement additional Tier 2 level structured intervention for students who did not respond adequately to the school-wide Tier 1 supports; this trial is referred to as the PBIS*plus* Trial (Bradshaw, Waasdorp, & Leaf, 2015). Data were collected on 29,569 students and 3,202 staff members across 42 Maryland elementary schools that were randomly assigned to either the Tier 1 only or the combined Tier 1 and Tier 2 intervention group. These data were collected at the fall of the school year (baseline, following the initial summer intervention), and the late springs of the current school year (Posttest 1) and two follow-up years (Posttests 2 and 3) on the TOCA-C.

Measures

The primary outcomes of interest in this study were the TOCA-C (Koth et al., 2009; Werthamer-Larsson et al., 1991) and the T-COMP (Conduct Problems Prevention Research Group, 2002). A benefit of these measures is that they are publicly available (free), commonly used measures of youth adjustment with adequate psychometric properties, thereby increasing the potential utility of the findings generated from this study. The TOCA-C is a checklist version of the original TOCA (TOCA-R; Werthamer-Larsson et al., 1991). For over 30 years, various versions of the TOCA have been used in large-scale RCTs to assess the impact of school-based prevention interventions (Bierman, et al., 2008; Conduct Problems Prevention Research Group,

2002; Ialongo et al., 1999; Kellam, Ling, Merisca, Brown, & Ialongo, 1998). Specifically, the TOCA-C (Koth et al., 2009) is a nonclinical measure of children's behavior completed by teachers. Subscales of the TOCA-C include concentration problems, disruptive behavior, prosocial behaviors, emotional dysregulation, internalization, family problems, and family involvement using a 6-point Likert-type scale (ranging from 1 = *never* to 6 = *almost always*). The scales were scored such that higher scores indicated better adjustment on the prosocial behavior and family involvement subscales, whereas lower scores indicate fewer symptoms (better adjustment) on the remaining subscales. These scales exhibit strong internal consistency (e.g., the Cronbach's αs range from .89 to .96; Bradshaw et al., 2015), have a consistent factor structure over time (Koth et al., 2009), relate to external criteria (Stormshak, Bierman, Bruschi, Dode, & Coie, 1999), and are sensitive to relatively modest intervention effects (Ialongo et al., 1999). Further, the TOCA has demonstrated high test–retest reliability (e.g., Werthamer-Larsson et al., 1991) and strong predictive validity (Petras et al., 2004). In fact, higher TOCA kindergarten scores were associated with more behavior problems at school, lower social skills, and poorer school adjustment reported by multiple informants (teacher, parent, and child) at the end of elementary, middle, and high school (Racz et al., 2013).

The T-COMP (Conduct Problems Prevention Research Group, 2002) is a measure of social and academic competence completed by teachers. This measure includes a 6-point Likert-type scale with item responses ranging from 0 (*almost never*) to 5 (*almost always*). There are three subscales (prosocial behavior, emotional regulation, and academic competence) and a total competence scale. Higher scores indicate higher levels of competence. Similar to the TOCA, the T-COMP has been widely used for large-scale RCTs to assess the impact of school-based prevention interventions (Bierman, et al., 2008; Conduct Problems Prevention Research Group, 2002; Webster-Stratton & Hammond, 1998). The T-COMP scales demonstrate strong internal consistency, have a consistent factor structure over time (Corrigan, 2003; Gifford-Smith, 2000), and are sensitive to intervention effects (Reinke et al., 2014; Webster-Stratton, 1998).

Overview of the Analyses

The current analytic sample included data from all four projects, based on the common outcome measures. Specifically, the analytic sample for the TOCA-C was combined from Projects 1, 3, and 4, which includes children from kindergarten to Grade 5; the analytic sample for the T-COMP was

combined from Projects 1 and 2, which includes children from kindergarten to Grade 2. The sample sizes of the analytic samples vary by the outcome measures and analyses (for calculating effect size benchmarks, ICCs) and proportions of variance explained by the covariates (R^2) due to missing data. Tables 1 and 2 present the descriptive characteristics by the outcome and grade for the samples to calculate ICC and R^2 using pretest measured at the beginning (fall) of same school year and using pretest measured at the end (spring) of previous school year, respectively. Applying similar approaches as Bloom et al. (2008) and Hill et al. (2008), we calculated two types of effect size benchmarks: (1) normative expectations for change and (2) policy-relevant demographic performance gaps. Methods consistent with the approach employed by Hedges and Hedberg (2007, 2013) were utilized in calculating ICCs and proportions of variance explained by the covariates (R^2) for two- and three-level cluster randomized experiments. Additional details of these calculations are provided below.

Benchmark 1: Normative expectations for change. The first empirical benchmark refers to expectations for annual growth or change in the absence of an intervention. In the context of prevention and educational sciences, the question is: "How does the effect of an intervention compare to a typical year of growth for a given target population of students?" (Hill et al., 2008, p. 173). Bloom et al. (2008) and Hill et al. (2008) calculated annual change in achievement in effect size by calculating the difference of mean scores in adjacent grades divided by pooled student-level standard deviation for the two adjacent grades. We applied SAS PROC MIXED (SAS Institute Inc., 2013) to fit the three-level HLM with students nested within classrooms, and teachers nested within schools to estimate the average annual change. The dependent variables were the simple score differences for the same children between two adjacent grades. The intercepts from the unconditional HLM estimated the average annual changes. We used the square root of the total variance (sum of Levels 1, 2, and 3 variances) to calculate the effect size of the annual change (Hedges, 2011; Spybrook, Hedges, & Borenstein, 2014; Spybrook & Raudenbush, 2009). Specifically, we used the average annual changes divided by the square roots of the average total variances of scores in adjacent grades to calculate the effect sizes. The 95% confidence intervals of the effect sizes were calculated based on the formulas in Hedges (2011). In addition to calculating annual changes using adjacent grades, we used measures at both the beginning and the end of grades. We calculated these two types of annual changes using data from the control group only.

Table 1. Descriptive Characteristics by Outcome and Grade for the Sample to Calculate ICCs and R^2 Using Pretest Measured at the Beginning (Fall) of Same School Year.

Outcome	Grade	Sample Size			White	Black	Hispanic	Female	Eligible for Free/Reduced Meals
		Student	Teacher	School					
TOCA-concentration problems	K	3,048	200	63	.33	.56	.06	.48	.42
TOCA-disruptive behavior	1	3,003	211	68	.35	.55	.06	.50	.49
TOCA-prosocial behavior	2	2,954	201	66	.34	.56	.06	.48	.49
	3	3,153	195	68	.34	.56	.06	.47	.47
	4	2,645	158	58	.36	.55	.05	.49	.43
	5	2,617	158	56	.35	.55	.06	.48	.42
TOCA-emotion dysregulation	K	2,635	162	48	.35	.55	.06	.48	.43
TOCA-internalization	1	2,557	158	50	.37	.54	.07	.51	.47
TOCA-family problems	2	2,430	150	49	.36	.55	.07	.48	.47
TOCA-family involvement	3	2,690	148	50	.36	.55	.06	.47	.46
	4	2,184	116	41	.38	.52	.06	.49	.40
	5	2,202	120	41	.37	.52	.07	.48	.40
COMP-prosocial behavior	K	832	65	20	.44	.51	.10	.49	.44
COMP-emotion regulation	1	705	54	15	.35	.57	.10	.50	.63
COMP-academic competence	2	752	54	18	.46	.49	.09	.48	.56
COMP-total scale									

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale; K = kindergarten. Entries for the characteristics are proportions.

Table 2. Descriptive Characteristics by Outcome and Grade for the Sample to Calculate ICCs and R^2 Using Pretest Measured at the End (Spring) of Previous School Year.

TOCA Outcomes	Grade	Sample Size			White	Black	Hispanic	Female	Eligible for Free/Reduced Meals
		Student	Teacher	School					
TOCA-concentration problems	1	4,178	329	97	.38	.50	.07	.48	.50
TOCA-disruptive behavior	2	5,832	437	133	.45	.44	.06	.48	.49
TOCA-prosocial behavior	3	5,967	436	133	.45	.46	.05	.48	.49
	4	4,715	322	102	.42	.48	.06	.49	.47
	5	3,294	208	66	.36	.53	.06	.49	.44
TOCA-emotion dysregulation	1	3,008	221	66	.35	.52	.08	.49	.50
TOCA-internalization	2	2,930	210	65	.37	.50	.08	.49	.48
TOCA-family problems	3	2,928	205	64	.38	.51	.08	.49	.46
TOCA-family involvement	4	3,141	201	65	.37	.51	.07	.48	.46
	5	3,294	208	66	.36	.53	.06	.49	.44

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Entries for the characteristics are proportions.

Benchmark 2: Demographic performance gaps among subgroups. It is well known that there are policy-relevant demographic academic performance gaps among subgroups, and educational policies have been made to reduce performance gaps regarding race/ethnicity, gender, and SES (Chatterji, 2006; Vanneman, Hamilton, Anderson, & Rahman, 2009). In this context, the question was: "How do the effects of an intervention compare with existing differences among subgroups of students?" (Hill et al., 2008, p. 174). We applied the three-level unconditional HLM to estimate the demographic performance gaps between two subgroups of children. The predictors of interest were the dummy variables indicating the demographic subgroups (White vs. Black, White vs. Hispanic, female vs. male, and Free and Reduced Meals (FARMs) eligible vs. ineligible), which were at student level and their slopes were allowed to randomly vary across classroom and school levels (we used the fixed slope models when the random slope models did not converge). The effect sizes were calculated by dividing the coefficients of the demographic subgroup indicators by the square roots of the total variances (sums of Levels 1, 2, and 3 variances; Hedges, 2011; Spybrook et al., 2014; Spybrook & Raudenbush, 2009). The 95% confidence intervals of the effect sizes were calculated based on the formulas in Hedges (2011). For the outcome measure at the beginning of each grade (i.e., pretest prior to receiving intervention), we used all participants from both study groups. For the posttest outcome measure after receiving the intervention, we used only the sample from the control group to estimate the natural gaps.

ICCs and proportions of variance explained by the covariates (R^2). Our outcomes were measured at the end of the school year, and we calculated the ICCs and the proportions of variance explained by the covariates (R^2) by grade (kindergarten to Grade 5). We used two types of pretests separately: measures at the end of the prior year (lower grade) and the beginning of the same year (same grade). The demographic variables included race, gender, and FARMs. Our original data have three levels, whereby students (Level 1) are nested within classrooms (Level 2; each teacher taught one classroom, hence we treat teachers same as classrooms), and classrooms nested within schools (Level 3), which allow us to calculate the ICCs and R^2 using three-level HLM for three-level CRTs. Our data also allow us to calculate the ICCs and R^2 using two-level HLM where students are nested within schools by ignoring classroom level for two-level CRTs. Although the data should be analyzed as the study was designed using three-level HLM, the additional analyses using two-level HLM will provide useful design parameters

for researchers to plan two-level CRTs when the classroom-level information is missing.

We used SAS PROC MIXED to estimate the variance components for estimating these parameters using four models: (1) unconditional model, (2) conditional on demographic variables, (3) conditional on pretest, and (4) conditional on demographic and pretest. Because the data that we used were from intervention studies, which were different from other design parameter studies that used data from observational studies without interventions (e.g., Hedges & Hedberg, 2007, 2013; Westine et al., 2013), we included the treatment condition in all the models to eliminate the treatment effects in estimating ICCs and R^2. In addition, because our data were combined from multiple projects in which the units for random assignment were schools or classrooms, the units for random assignment were schools in the combined data for Grades 3, 4, and 5, and the treatment status appeared to vary across classrooms within schools for kindergarten to Grade 2. Hence, we put the treatment status at Level 3 (schools) for the analyses of the outcomes in Grades 3, 4, and 5, and we put the treatment status at Level 2 (classrooms) for the analyses of the outcomes in kindergarten and Grades 1 and 2 and assumed the treatment effect constant across schools. We made the decision to use the fixed slope model for the analyses of the outcomes in Kindergarten and Grades 1 and 2 because of the empirical consideration of the complicated data structure. In the combined data for the analyses of the outcomes in kindergarten and Grades 1 and 2, there were only 9 schools in which classrooms were randomly assigned to the treatment groups, and there were many other schools (6–124) in which schools were randomly assigned to the treatment groups that makes it impossible to estimate treatment effect variability across schools (see sample sizes in Tables 1 and 2). However, the treatment effect variation is an important topic, and it should not be ignored when the data are available (for more discussion about the random slope model, see Aguinis & Culpepper, 2015; Berkhof & Kampen, 2004). Our approach would estimate the appropriate ICCs and R^2 for power analyses in assisting designing studies in which either students, classrooms, or schools are the units for random assignment because they are the common design parameters no matter which level the units for random assignment is at. We present the first two three-level HLMs using the notation employed by Raudenbush and Bryk (2002) below and put the other HLMs in the supplement to save space.

The "unconditional" model. The unconditional model is the one that is unconditional on any covariates, but still controls for the treatment status to

eliminate any potential effects on the variance components from the intervention. The three-level unconditional model for the analyses of the outcomes in kindergarten and Grades 1 and 2 is:

$$\text{Level1 (student): } Y_{ijk} = \alpha_{0jk} + e_{ijk}, e_{ijk} \sim N(0, \sigma^2)$$

$$\text{Level2 (class): } \quad \alpha_{0jk} = \beta_{00k} + \beta_{01k}(\text{Condition})_{jk} + u_{jk}, u_{jk} \sim N(0, \tau_2^2)$$

$$\text{Level3 (school): } \begin{matrix} \beta_{00k} = \gamma_{000} + \xi_k \\ \beta_{01k} = \gamma_{001} \end{matrix}, \xi \sim N(0, \tau_3^2)$$

where Y_{ijk} is the outcome variable for student i in class j in school k; $(\text{Condition})_{jk}$ is a binary variable indicating treatment condition (Condition = 0 for control group and Condition = 1 for treatment group). Note that $(\text{Condition})_k$ is at Level 3 for the analyses of the outcomes in Grades 3, 4, and 5. Our primary interest was in estimating the variance components of Level-1 (σ^2), Level-2 (τ_2^2), and Level-3 (τ_3^2). The unconditional ICCs for teacher and school level are $\rho_2 = \tau_2^2/(\sigma^2 + \tau_2^2 + \tau_3^2)$ and $\rho_3 = \tau_3^2/(\sigma^2 + \tau_2^2 + \tau_3^2)$, respectively. According to Hedges, Hedberg, and Kuyper (2012), the standard errors of ρ_2 and ρ_3 in the large sample balanced three-level models are $SE(\rho_2) = \sqrt{[J(1-\rho_2)^2 + 2\rho_2(1-\rho_2)]v_2 + J\rho_2^2 v_3 / J(\sigma^2 + \tau_2^2 + \tau_3^2)^2}$ and $SE(\rho_3) = \sqrt{[J\rho_3^2 + 2\rho_3(1-\rho_3)]v_2 + J(1-\rho_3)^2 v_3 / J(\sigma^2 + \tau_2^2 + \tau_3^2)^2}$, respectively, where v_2 and v_3 are variances components of τ_2^2 and τ_3^2, respectively, and J is the harmonic mean number of teachers per school.

The pretest covariate model. We used the pretest group mean centering as used by Hedges and Hedberg (2007, 2013) because it leads to more stable estimates of variance components. For the pretest (X_{ijk}) of child i in class j in school k, we first calculate the class-means $(\bar{X}_{\bullet jk})$ of pretest for class j in school k, then the class-centered pretest is $(X_{ijk} - \bar{X}_{\bullet jk})$ that is used as Level-1 covariate. Second, we calculate the school-means by $\bar{X}_{\bullet\bullet k} = (1/K)\sum_{k=1}^{K} \bar{X}_{\bullet jk}$ that is used as Level-3 covariate, and the school-centered pretest is $(\bar{X}_{\bullet jk} - \bar{X}_{\bullet\bullet k})$ that is used the Level-2 covariate. By decomposing the students' pretest into class-centered pretest (Level 1), school-centered pretest (Level 2), and school-means (Level 3), and putting them at each level, a three-level HLM allows us to estimate three coefficients for three-level pretest. This group-mean centering approach provides more accurate estimates of the effects of pretest on posttest than without

group-mean centering which only estimates one coefficient for the students' pretest. Hence, it may improve precision of estimates of variances components. One might use grand-mean centering before using the group-mean centering, but it will not change the estimates of treatment effects or variance components although it will change the interpretation of the intercept which was not of interest in the intervention studies. The detailed model for the analyses of the outcomes in Kindergarten and Grades 1 and 2 is as follows (Similarly, $(\text{Condition})_k$ is at Level 3 for the analyses of the outcomes in Grades 3, 4, and 5.):

$$
\begin{aligned}
\text{Level 1(student):} \quad & Y_{ijk} = \alpha_{0jk} + \alpha_{1jk}(X_{ijk} - \bar{X}_{\bullet jk}) + e_{ijk}, e_{ijk} \sim N(0, \sigma^2_{|X}) \\
\text{Level 2(class):} \quad & \alpha_{0jk} = \beta_{00k} + \beta_{01k}(\text{Condition})_{jk} \\
& \qquad +\beta_{02k}(\bar{X}_{\bullet jk} - \bar{X}_{\bullet\bullet k}) + \mu_{jk}, u_{jk} \sim N(0, \tau^2_{2|X}) \\
& \alpha_{1jk} = \beta_{10k} \\
\text{Level 3(school):} \quad & \beta_{00k} = \gamma_{000} + \gamma_{001}\bar{X}_{\bullet\bullet k} + \xi_k \\
& \beta_{01k} = \gamma_{001} \\
& \beta_{02k} = \gamma_{002} \qquad\qquad , \xi \sim N(0, \tau^2_{3|X}) \\
& \beta_{10k} = \gamma_{100}
\end{aligned}
$$

where $\sigma^2_{|X}$, $\tau^2_{2|X}$, and $\tau^2_{3|X}$ are variance components for Level 1, Level 2, and Level 3 conditional on pretest, respectively. The pseudo-R^2 for pretest at Level 1, Level 2, and Level 3 are $R^2_1 = (1 - \sigma^2_{|X})/\sigma^2$, $R^2_2 = (1 - \tau^2_{2|X})/\tau^2_2$, and $R^2_3 = (1 - \tau^2_{3|X})/\tau^2_3$, respectively. These pseudo-R^2 indicate the proportions of variance explained by pretest in addition to the treatment variable, which can be used to calculate the standard error of the treatment effect estimate and statistical power.

The demographic covariates model. We did not use demographic covariates group-mean centering because we found that it did not improve the model fit statistics as it did in the pretest covariate model. Rather, we included demographic covariates (race, gender, and FARMs) at Level 1. The detailed model is in Appendix A.

The pretest and demographic covariates model. This model including the pretest with group-mean centering and the demographic covariates without group-mean centering is in Appendix A.

Note that in all the four models, we estimated the unconditional ICCs and R^2 after controlling for the treatment effects. This approach allows us to estimate the appropriate ICCs and R^2 for power analyses using data from

intervention studies. We also estimated the unconditional ICCs and R^2 only using the control sample. When researchers conduct data analyses of CRTs using HLM, the treatment variable should be put at the level at which the units are randomized to the treatment and control groups.

To summarize our findings about the unconditional ICCs and pseudo-R^2, we calculated the mean ICCs and R^2 across "outcome measures" and grades, and the Pearson correlations of ICCs and R^2 with grade across outcome measures (Hedges & Hedberg, 2007). We also fit the regression models by regressing ICCs and R^2 on the outcome measures and grades. The F-statistic of the predictor outcome measures and the regression coefficients of the predictor grade were reported.

Results

Normative Expectations for Change

The results of the normative expectations for change as measured by the difference between the beginning (fall) and the end (spring) of same school year are presented in Table 3. Some TOCA-C subscales (e.g., concentration problems, prosocial behavior, and family problems) illustrated small gains (<.10) and inconsistent significance[1] patterns and directional (better or worse) patterns across grades. Other TOCA-C subscales (e.g., disruptive behavior, emotion dysregulation, and internalizing) had gains ranging from .08 to .26 across grades, and all were significantly worse at the end of school year compared with beginning of the school year. The TOCA-C family involvement scale illustrated the same significantly worse change from the beginning to the end of school year for Kindergarten (Effect Size (ES) = −.11) and Grade 5 students (ES = −.14), but small and insignificant change (ranging from −.02 to .05) for students in Grades 1–4. The T-COMP prosocial behavior had small and insignificant gains for Kindergarten (ES = .07) and Grade 1 (ES = .06) students, but large and significant gains for Grade 2 (ES = .18) students. The other T-COMP subscales (e.g., emotion regulation, academic competence, and the COMP total scale) had consistently significant positive changes (ranging from .27 to .38) from the beginning to the end of school year for children in Kindergarten through Grade 2.

The results of the normative expectations for change on the TOCA-C outcomes as measured by the difference between two adjacent years are presented in Table C1. In general, the annual changes are small and insignificant. Most gains were less than .10. However, concentration problems were significantly worse for children from kindergarten to Grade 1 (ES = .18); family

Table 3. Annual Changes from the Beginning (Fall) to the End (Spring) of School Year.

Outcome	Effect Size	Low	High	Student	Teacher	School
Kindergarten						
TOCA-concentration problems	−.03	−.09	.03	2,393	137	48
TOCA-disruptive behavior	.15	.08	.21	2,393	137	48
TOCA-prosocial behavior	−.01	−.07	.05	2,393	137	48
TOCA-emotion dysregulation	.13	.03	.23	1,578	95	32
TOCA-internalization	.14	.01	.26	1,578	95	32
TOCA-family problems	−.06	−.28	.16	1,578	95	32
TOCA-family involvement	−.11	−.23	.01	1,381	84	26
COMP-prosocial behavior	.07	−.11	.24	501	37	14
COMP-emotion regulation	.27	−.12	.67	501	37	14
COMP-academic competence	.38	.15	.60	501	37	14
COMP-total scale	.30	.09	.50	501	37	14
Grade 1						
TOCA-concentration problems	−.08	−.17	.02	2,383	147	49
TOCA-disruptive behavior	.19	.13	.24	2,383	147	49
TOCA-prosocial behavior	.01	−.06	.07	2,383	147	49
TOCA-emotion dysregulation	.09	.03	.15	1,427	89	33
TOCA-internalization	.16	.06	.25	1,427	89	33
TOCA-family problems	−.10	−.35	.15	1,427	89	33
TOCA-family involvement	.03	−.08	.14	1,306	82	29
COMP-prosocial behavior	.06	−.09	.20	344	25	12
COMP-emotion regulation	.37	−.14	.88	344	25	12
COMP-academic competence	.27	.12	.42	344	25	12
COMP-total scale	.27	.05	.49	344	25	12
Grade 2						
TOCA-concentration problems	−.09	−.20	.01	2,416	148	48
TOCA-disruptive behavior	.21	.14	.27	2,416	148	48
TOCA-prosocial behavior	.03	−.04	.11	2,416	148	48
TOCA-emotion dysregulation	.13	.04	.23	1,413	88	32
TOCA-internalization	.17	.04	.29	1,413	88	32
TOCA-family problems	−.12	−.36	.12	1,413	88	32
TOCA-family involvement	.04	−.06	.15	1,265	80	28
COMP-prosocial behavior	.18	−.02	.37	365	26	13
COMP-emotion regulation	.36	−.03	.76	365	26	13
COMP-academic competence	.27	.17	.36	365	26	13
COMP-total scale	.32	.16	.48	365	26	13
Grade 3						
TOCA-concentration problems	.04	.00	.09	2,546	139	46
TOCA-disruptive behavior	.23	.17	.30	2,546	139	46
TOCA-prosocial behavior	−.08	−.15	−.02	2,546	139	46

(continued)

Table 3. (continued)

Outcome	Effect Size	Low	High	Student	Teacher	School
TOCA-emotion dysregulation	.13	.04	.22	1,526	83	30
TOCA-internalization	.10	.03	.18	1,526	83	30
TOCA-family problems	.02	−.08	.12	1,526	83	30
TOCA-family involvement	−.02	−.15	.12	1,469	79	27
Grade 4						
TOCA-concentration problems	.00	−.05	.05	2,369	123	39
TOCA-disruptive behavior	.20	.14	.27	2,369	123	39
TOCA-prosocial behavior	−.03	−.09	.04	2,369	123	39
TOCA-emotion dysregulation	.08	.02	.15	1,345	71	23
TOCA-internalization	.13	.03	.23	1,345	71	23
TOCA-family problems	.05	−.07	.16	1,345	71	23
TOCA-family involvement	.05	−.07	.16	1,345	71	23
Grade 5						
TOCA-concentration problems	.08	.03	.13	2,415	122	39
TOCA-disruptive behavior	.26	.20	.32	2,415	122	39
TOCA-prosocial behavior	−.08	−.16	−.01	2,415	122	39
TOCA-emotion dysregulation	.14	.06	.21	1,324	68	23
TOCA-internalization	.16	.05	.27	1,324	68	23
TOCA-family problems	.12	.00	.24	1,324	68	23
TOCA-family involvement	−.14	−.32	.05	1,324	68	23

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Low and high refer to the lower and higher bounds of the 95% confidence intervals of effect sizes, respectively. Student, Teacher, and School refer to the number of students, teachers, and schools for the data analyses, respectively. Higher scores indicate more favorable on the TOCA-C prosocial behavior and family involvement subscales, and all T-COMP subscales. Lower scores on TOCA-C concentration, disruptive behavior, emotion dysregulation, internalization, and family problems subscales indicate more favorable adjustment.

involvement was significantly worse for children from kindergarten to Grade 1 (ES = −.23), but significantly better from Grades 1 to 2 (ES = .13). In summary, the TOCA-C measures did not illustrate consistent patterns on yearly changes, but most T-COMP measures illustrated significant changes from the beginning to the end of school years and same directional results.

Demographic Performance Gaps Among Subgroups

Tables 4 through 7 report standardized mean difference effect sizes for the performance differences between selected subgroups on the TOCA-C and

Table 4. White–African American Gaps at the End (Spring) of School Year.

Outcome	Effect Size	Low	High	Student	Teacher	School
Kindergarten						
TOCA-concentration problems	−.27	−.32	−0.22	9,110	480	146
TOCA-disruptive behavior	−.27	−.32	−0.22	9,110	480	146
TOCA-prosocial behavior	.18	.13	0.22	9,110	480	146
TOCA-emotion dysregulation	−.23	−.30	−0.15	4,778	264	76
TOCA-internalization	−.05	−.12	0.03	4,778	264	76
TOCA-family problems	−.13	−.21	−0.04	4,778	264	76
TOCA-family involvement	.39	.31	0.48	4,778	264	76
COMP-prosocial behavior	.23	.03	0.43	501	37	14
COMP-emotion regulation	.15	−.11	0.42	501	37	14
COMP-academic competence	.36	.16	0.57	501	37	14
COMP-total scale	.29	.09	0.49	501	37	14
Grade 1						
TOCA-concentration problems	−.24	−.29	−0.19	11,328	648	188
TOCA-disruptive behavior	−.33	−.38	−0.28	11,328	648	188
TOCA-prosocial behavior	.18	.14	0.23	11,328	648	188
TOCA-emotion dysregulation	−.25	−.33	−0.17	4,506	254	77
TOCA-internalization	−.04	−.11	0.03	4,506	254	77
TOCA-family problems	−.12	−.21	−0.03	4,506	254	77
TOCA-family involvement	.36	.28	0.43	4,506	254	77
COMP-prosocial behavior	.39	.12	0.66	344	25	12
COMP-emotion regulation	.24	−.06	0.53	344	25	12
COMP-academic competence	.57	.20	0.94	344	25	12
COMP-total scale	.46	.20	0.73	344	25	12
Grade 2						
TOCA-concentration problems	−.23	−.27	−0.19	13,517	781	223
TOCA-disruptive behavior	−.34	−.38	−0.29	13,517	781	223
TOCA-prosocial behavior	.22	.18	0.27	13,517	781	223
TOCA-emotion dysregulation	−.19	−.28	−0.11	4,589	260	76
TOCA-internalization	−.01	−.07	0.06	4,589	260	76
TOCA-family problems	−.08	−.16	0.00	4,589	260	76
TOCA-family involvement	.35	.27	0.44	4,589	260	76
COMP-prosocial behavior	.36	.05	0.67	365	26	13
COMP-emotion regulation	.34	.06	0.62	365	26	13
COMP-academic competence	.63	.26	1.00	365	26	13
COMP-total scale	.54	.18	0.91	365	26	13
Grade 3						
TOCA-concentration problems	−.26	−.30	−0.21	13,376	735	221
TOCA-disruptive behavior	−.39	−.43	−0.34	13,376	735	221

(continued)

Table 4. (continued)

Outcome	Effect Size	Low	High	Student	Teacher	School
TOCA-prosocial behavior	.22	.18	0.26	13,376	735	221
TOCA-emotion dysregulation	−.16	−.25	−0.08	4,469	232	74
TOCA-internalization	.05	−.03	0.14	4,469	232	74
TOCA-family problems	−.13	−.21	−0.06	4,469	232	74
TOCA-family involvement	.41	.32	0.49	4,469	232	74
Grade 4						
TOCA-concentration problems	−.28	−.33	−0.23	11,316	597	178
TOCA-disruptive behavior	−.41	−.47	−0.36	11,316	597	178
TOCA-prosocial behavior	.24	.19	0.29	11,316	597	178
TOCA-emotion dysregulation	−.22	−.30	−0.15	4,727	232	67
TOCA-internalization	.02	−.05	0.09	4,727	232	67
TOCA-family problems	−.04	−.11	0.03	4,727	232	67
TOCA-family involvement	.38	.29	0.47	4,727	232	67
Grade 5						
TOCA-concentration problems	−.31	−.37	−0.26	9,233	469	141
TOCA-disruptive behavior	−.40	−.46	−0.35	9,233	469	141
TOCA-prosocial behavior	.18	.14	0.22	9,233	469	141
TOCA-emotion dysregulation	−.31	−.38	−0.23	4,707	229	67
TOCA-internalization	−.01	−.08	0.07	4,707	229	67
TOCA-family problems	−.08	−.16	0.00	4,707	229	67
TOCA-family involvement	.36	.29	0.43	4,707	229	67

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Low and high refer to the lower and higher bounds of the 95% confidence intervals of effect sizes, respectively. Student, Teacher, and School refer to the number of students, teachers, and schools for the data analyses, respectively. Higher scores indicate more favorable on the TOCA-C prosocial behavior and family involvement subscales, and all T-COMP subscales. Lower scores on TOCA-C concentration, disruptive behavior, emotion dysregulation, internalization, and family problems subscales indicate more favorable adjustment.

T-COMP measures. Statistically significant large gaps existed for most measures across subgroups favoring students who were White, female, and ineligible for FARMs. With the exception of two TOCA-C measures (internalization and family problems), the difference between Whites and African Americans, the TOCA-C and T-COMP measures illustrated significant gaps (the absolute values of ES ranging from .16 to .63) favoring Whites, which were consistent across grades (Table 4). Regarding the difference between Whites and Hispanic, most measures did not have consistent patterns (Table 5). However, compared to Hispanics, Whites had significantly higher scores on the TOCA-C family involvement measure from

Table 5. White–Hispanic Gaps at the End (Spring) of School Year.

Outcome	Effect Size	Low	High	Student	Teacher	School
Kindergarten						
TOCA-concentration problems	.07	−.02	0.16	9,110	480	146
TOCA-disruptive behavior	.11	.01	0.21	9,110	480	146
TOCA-prosocial behavior	.03	−.04	0.10	9,110	480	146
TOCA-emotion dysregulation	.26	.14	0.38	4,778	264	76
TOCA-internalization	.07	−.04	0.17	4,778	264	76
TOCA-family problems	.14	.02	0.25	4,778	264	76
TOCA-family involvement	.44	.31	0.56	4,778	264	76
COMP-prosocial behavior	−.28	−.55	−0.01	501	37	14
COMP-emotion regulation	−.35	−.67	−0.03	501	37	14
COMP-academic competence	.14	−.13	0.41	501	37	14
COMP-total scale	−.09	−.35	0.18	501	37	14
Grade 1						
TOCA-concentration problems	−.05	−.14	0.04	11,328	648	188
TOCA-disruptive behavior	.04	−.05	0.14	11,328	648	188
TOCA-prosocial behavior	−.04	−.11	0.03	11,328	648	188
TOCA-emotion dysregulation	.20	.06	0.33	4,506	254	77
TOCA-internalization	.18	.06	0.29	4,506	254	77
TOCA-family problems	.12	−.01	0.26	4,506	254	77
TOCA-family involvement	.31	.19	0.43	4,506	254	77
COMP-prosocial behavior	.21	−.21	0.63	344	25	12
COMP-emotion regulation	−.26	−.59	0.07	344	25	12
COMP-academic competence	.83	.33	1.34	344	25	12
COMP-total scale	.48	.09	0.87	344	25	12
Grade 2						
TOCA-concentration problems	.07	−.02	0.15	13,517	781	223
TOCA-disruptive behavior	.11	.02	0.19	13,517	781	223
TOCA-prosocial behavior	−.05	−.12	0.01	13,517	781	223
TOCA-emotion dysregulation	.20	.07	0.32	4,589	260	76
TOCA-internalization	.15	.05	0.26	4,589	260	76
TOCA-family problems	.00	−.12	0.11	4,589	260	76
TOCA-family involvement	.36	.24	0.47	4,589	260	76
COMP-prosocial behavior	.06	−.43	0.56	365	26	13
COMP-emotion regulation	.33	−.10	0.76	365	26	13
COMP-academic competence	.81	.42	1.20	365	26	13
COMP-total scale	.50	.06	0.95	365	26	13
Grade 3						
TOCA-concentration problems	.02	−.06	0.10	13,376	735	221
TOCA-disruptive behavior	.07	−.01	0.14	13,376	735	221

(continued)

Table 5. (continued)

Outcome	Effect Size	Low	High	Student	Teacher	School
TOCA-prosocial behavior	−.04	−.11	0.03	13,376	735	221
TOCA-emotion dysregulation	.17	.04	0.31	4,469	232	74
TOCA-internalization	.10	−.02	0.22	4,469	232	74
TOCA-family problems	.03	−.11	0.17	4,469	232	74
TOCA-family involvement	.41	.30	0.53	4,469	232	74
Grade 4						
TOCA-concentration problems	.03	−.06	0.11	11,316	597	178
TOCA-disruptive behavior	.09	.00	0.18	11,316	597	178
TOCA-prosocial behavior	−.03	−.10	0.05	11,316	597	178
TOCA-emotion dysregulation	.24	.12	0.36	4,727	232	67
TOCA-internalization	.22	.11	0.34	4,727	232	67
TOCA-family problems	.14	.02	0.26	4,727	232	67
TOCA-family involvement	.36	.24	0.48	4,727	232	67
Grade 5						
TOCA-concentration problems	−.02	−.13	0.09	9,233	469	141
TOCA-disruptive behavior	.08	−.02	0.17	9,233	469	141
TOCA-prosocial behavior	−.03	−.12	0.05	9,233	469	141
TOCA-emotion dysregulation	.11	−.01	0.23	4,707	229	67
TOCA-internalization	.19	.07	0.30	4,707	229	67
TOCA-family problems	.08	−.04	0.20	4,707	229	67
TOCA-family involvement	.39	.29	0.50	4,707	229	67

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Low and high refer to the lower and higher bounds of the 95% confidence intervals of effect sizes, respectively. Student, Teacher, and School refer to the number of students, teachers, and schools for the data analyses, respectively. Higher scores indicate more favorable on the TOCA-C prosocial behavior and family involvement subscales, and all T-COMP subscales. Lower scores on TOCA-C concentration, disruptive behavior, emotion dysregulation, internalization, and family problems subscales indicate more favorable adjustment.

kindergarten to Grade 5 (ES ranging from .31 to .44), while Hispanics had significantly more adaptive scores than Whites on the TOCA-C emotion dysregulation measure from kindergarten to Grade 5 (ES ranging from .11 to .26). Regarding the difference between girls and boys, the TOCA-C and T-COMP measures, with the exception of the TOCA-C family involvement measure, illustrated significant gaps (the absolute values of ES ranging from .08 to .54) favoring girls, which were consistent across grades (Table 6). Regarding the difference between the FARMs-eligible and FARMs-ineligible children, the TOCA-C and T-COMP measures, except the T-COMP emotion dysregulation measure, illustrated significant gaps

Table 6. Female–Male Gaps at the End (Spring) of School Year.

Outcome	Effect Size	Low	High	Student	Teacher	School
Kindergarten						
TOCA-concentration problems	−.44	−.49	−.40	9,110	480	146
TOCA-disruptive behavior	−.31	−.34	−.27	9,110	480	146
TOCA-prosocial behavior	.19	.16	.22	9,110	480	146
TOCA-emotion dysregulation	−.37	−.43	−.32	4,778	264	76
TOCA-internalization	−.10	−.14	−.05	4,778	264	76
TOCA-family problems	−.11	−.16	−.06	4,778	264	76
TOCA-family involvement	.03	−.01	.08	4,778	264	76
COMP-prosocial behavior	.45	.30	.60	501	37	14
COMP-emotion regulation	.47	.28	.66	501	37	14
COMP-academic competence	.44	.25	.63	501	37	14
COMP-total scale	.51	.33	.69	501	37	14
Grade 1						
TOCA-concentration problems	−.43	−.47	−.40	11,328	648	188
TOCA-disruptive behavior	−.33	−.37	−.30	11,328	648	188
TOCA-prosocial behavior	.20	.17	.23	11,328	648	188
TOCA-emotion dysregulation	−.40	−.45	−.35	4,506	254	77
TOCA-internalization	−.12	−.17	−.08	4,506	254	77
TOCA-family problems	−.10	−.15	−.04	4,506	254	77
TOCA-family involvement	.08	.03	.13	4,506	254	77
COMP-prosocial behavior	.54	.35	.74	344	25	12
COMP-emotion regulation	.39	.18	.60	344	25	12
COMP-academic competence	.33	.01	.65	344	25	12
COMP-total scale	.44	.25	.64	344	25	12
Grade 2						
TOCA-concentration problems	−.51	−.54	−.47	13,517	781	223
TOCA-disruptive behavior	−.38	−.42	−.35	13,517	781	223
TOCA-prosocial behavior	.28	.25	.31	13,517	781	223
TOCA-emotion dysregulation	−.43	−.49	−.37	4,589	260	76
TOCA-internalization	−.11	−.16	−.06	4,589	260	76
TOCA-family problems	−.12	−.19	−.05	4,589	260	76
TOCA-family involvement	.09	.04	.15	4,589	260	76
COMP-prosocial behavior	.43	.19	.67	365	26	13
COMP-emotion regulation	.46	.22	.70	365	26	13
COMP-academic competence	.24	−.04	.53	365	26	13
COMP-total scale	.42	.14	.69	365	26	13
Grade 3						
TOCA-concentration problems	−.52	−.55	−.49	13,376	735	221
TOCA-disruptive behavior	−.41	−.44	−.37	13,376	735	221

(continued)

Table 6. (continued)

Outcome	Effect Size	Low	High	Student	Teacher	School
TOCA-prosocial behavior	.29	.26	.32	13,376	735	221
TOCA-emotion dysregulation	−.45	−.51	−.38	4,469	232	74
TOCA-internalization	−.14	−.20	−.08	4,469	232	74
TOCA-family problems	−.14	−.21	−.07	4,469	232	74
TOCA-family involvement	.06	−.01	.12	4,469	232	74
Grade 4						
TOCA-concentration problems	−.53	−.56	−.49	11,316	597	178
TOCA-disruptive behavior	−.41	−.44	−.37	11,316	597	178
TOCA-prosocial behavior	.28	.25	.31	11,316	597	178
TOCA-emotion dysregulation	−.47	−.53	−.41	4,727	232	67
TOCA-internalization	−.09	−.14	−.04	4,727	232	67
TOCA-family problems	−.07	−.12	−.01	4,727	232	67
TOCA-family involvement	.08	.03	.13	4,727	232	67
Grade 5						
TOCA-concentration problems	−.53	−.57	−.48	9,233	469	141
TOCA-disruptive behavior	−.42	−.46	−.37	9,233	469	141
TOCA-prosocial behavior	.29	.24	.33	9,233	469	141
TOCA-emotion dysregulation	−.40	−.46	−.34	4,707	229	67
TOCA-internalization	−.08	−.14	−.03	4,707	229	67
TOCA-family problems	−.08	−.13	−.02	4,707	229	67
TOCA-family involvement	.03	−.03	.08	4,707	229	67

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Low and high refer to the lower and higher bounds of the 95% confidence intervals of effect sizes, respectively. Student, Teacher, and School refer to the number of students, teachers, and schools for the data analyses, respectively. Higher scores indicate more favorable on the TOCA-C prosocial behavior and family involvement subscales, and all T-COMP subscales. Lower scores on TOCA-C concentration, disruptive behavior, emotion dysregulation, internalization, and family problems subscales indicate more favorable adjustment.

(the absolute values of ES ranging from .07 to .63) favoring the FARMs-ineligible children, which are consistent across grades (Table 7).

ICCs and Proportions of Variance Explained by the Covariates (R^2)

The point estimates of the unconditional ICCs and R^2 by using the full sample and using the control sample are similar. We report the unconditional ICCs at the school level (ρ) for two-level HLM, ICCs at the school level (ρ_3) and class level (ρ_2) for a three-level HLM, and their 95% confidence intervals by grade as well as the summary statistics in Tables 8 and

Table 7. Eligible–Ineligible Free or Reduced-Price Meals Gaps at the End (Spring) of School Year.

Outcome	Effect Size	Low	High	Student	Teacher	School
Kindergarten						
TOCA-concentration problems	.31	.26	.37	9,110	480	146
TOCA-disruptive behavior	.17	.12	.23	9,110	480	146
TOCA-prosocial behavior	−.13	−.18	−.09	9,110	480	146
TOCA-emotion dysregulation	.13	.07	.20	4,778	264	76
TOCA-internalization	.06	−.01	.12	4,778	264	76
TOCA-family problems	.31	.23	.40	4,778	264	76
TOCA-family involvement	−.53	−.62	−.44	4,778	264	76
COMP-prosocial behavior	−.28	−.45	−.10	501	37	14
COMP-emotion regulation	−.16	−.34	.01	501	37	14
COMP-academic competence	−.51	−.69	−.33	501	37	14
COMP-total scale	−.39	−.54	−.25	501	37	14
Grade 1						
TOCA-concentration problems	.31	.26	.36	11,328	648	188
TOCA-disruptive behavior	.21	.15	.26	11,328	648	188
TOCA-prosocial behavior	−.15	−.20	−.11	11,328	648	188
TOCA-emotion dysregulation	.14	.08	.21	4,506	254	77
TOCA-internalization	.10	.03	.16	4,506	254	77
TOCA-family problems	.34	.26	.42	4,506	254	77
TOCA-family involvement	−.58	−.68	−.48	4,506	254	77
COMP-prosocial behavior	−.20	−.43	.04	344	25	12
COMP-emotion regulation	−.18	−.42	.06	344	25	12
COMP-academic competence	−.32	−.55	−.09	344	25	12
COMP-total scale	−.30	−.53	−.07	344	25	12
Grade 2						
TOCA-concentration problems	.34	.29	.38	13,517	781	223
TOCA-disruptive behavior	.21	.16	.26	13,517	781	223
TOCA-prosocial behavior	−.18	−.23	−.14	13,517	781	223
TOCA-emotion dysregulation	.12	.05	.19	4,589	260	76
TOCA-internalization	.07	.02	.13	4,589	260	76
TOCA-family problems	.34	.25	.43	4,589	260	76
TOCA-family involvement	−.55	−.63	−.48	4,589	260	76
COMP-prosocial behavior	−.15	−.44	.14	365	26	13
COMP-emotion regulation	−.01	−.24	.22	365	26	13
COMP-academic competence	−.38	−.79	.03	365	26	13
COMP-total scale	−.25	−.62	.12	365	26	13
Grade 3						
TOCA-concentration problems	.34	.29	.38	13,376	735	221
TOCA-disruptive behavior	.23	.19	.27	13,376	735	221

(continued)

Table 7. (continued)

Outcome	Effect Size	Low	High	Student	Teacher	School
TOCA-prosocial behavior	−.18	−.22	−.14	13,376	735	221
TOCA-emotion dysregulation	.16	.09	.23	4,469	232	74
TOCA-internalization	.12	.05	.18	4,469	232	74
TOCA-family problems	.32	.22	.41	4,469	232	74
TOCA-family involvement	−.60	−.67	−.52	4,469	232	74
Grade 4						
TOCA-concentration problems	.33	.28	.38	11,316	597	178
TOCA-disruptive behavior	.24	.19	.29	11,316	597	178
TOCA-prosocial behavior	−.16	−.20	−.12	11,316	597	178
TOCA-emotion dysregulation	.23	.15	.31	4,727	232	67
TOCA-internalization	.17	.10	.24	4,727	232	67
TOCA-family problems	.39	.30	.48	4,727	232	67
TOCA-family involvement	−.56	−.65	−.46	4,727	232	67
Grade 5						
TOCA-concentration problems	.32	.26	.37	9,233	469	141
TOCA-disruptive behavior	.26	.20	.31	9,233	469	141
TOCA-prosocial behavior	−.13	−.17	−.09	9,233	469	141
TOCA-emotion dysregulation	.21	.13	.28	4,707	229	67
TOCA-internalization	.07	.01	.12	4,707	229	67
TOCA-family problems	.33	.25	.41	4,707	229	67
TOCA-family involvement	−.49	−.56	−.42	4,707	229	67

Note. TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale. Low and high refer to the lower and higher bounds of the 95% confidence intervals of effect sizes, respectively. Student, Teacher, and School refer to the number of students, teachers, and schools for the data analyses, respectively. Higher scores indicate more favorable on the TOCA-C prosocial behavior and family involvement subscales, and all T-COMP subscales. Lower scores on TOCA-C concentration, disruptive behavior, emotion dysregulation, internalization, and family problems subscales indicate more favorable adjustment.

C2 for the full sample described in Tables 1 and 2, respectively. For two-level HLM, the TOCA-C prosocial behavior measure had a school-level ICC ranging from .29 to .54 while all the other measures had a school-level ICC ranging from .03 to .23. All the ICCs and their lower bounds of the 95% confidence intervals were not negative in two-level HLM, which indicated that there was significant between-school variance. For three-level HLM, as expected, the school-level ICC (ρ_3) is smaller than the school-level ICC (ρ) in two-level HLM because the class-level variances were allocated to school and student levels when two-level HLM was used. The TOCA-C prosocial behavior measure had larger school-level ICCs (ρ_3 ranging from

Table 8. Unconditional ICCs for a Two- and Three-Level HLM by Grade: Pretest Measured at the Beginning (Fall) of Same School Year.

	Two-Level HLM			Three-Level HLM					
Outcome	ρ	Low	High	ρ_3	Low	High	ρ_2	Low	High
Kindergarten									
TOCA-concentration problems	0.08	0.04	0.11	0.04	0.01	0.08	0.10	0.06	0.13
TOCA-disruptive behavior	0.12	0.07	0.17	0.06	0.01	0.11	0.17	0.12	0.22
TOCA-prosocial behavior	0.54	0.45	0.63	0.49	0.39	0.60	0.13	0.08	0.17
TOCA-emotion dysregulation	0.09	0.05	0.13	0.03	−0.01	0.08	0.20	0.14	0.25
TOCA-internalization	0.19	0.12	0.26	0.07	0.00	0.15	0.31	0.23	0.39
TOCA-family problems	0.20	0.13	0.27	0.16	0.08	0.24	0.13	0.09	0.18
TOCA-family involvement	0.14	0.09	0.20	0.09	0.03	0.15	0.18	0.13	0.23
COMP-prosocial behavior	0.15	0.05	0.25	0.07	−0.03	0.17	0.19	0.09	0.30
COMP-emotion regulation	0.10	0.02	0.18	0.02	−0.06	0.09	0.24	0.13	0.35
COMP-academic competence	0.14	0.05	0.22	0.11	0.02	0.20	0.08	0.02	0.14
COMP-total scale	0.16	0.06	0.26	0.10	−0.01	0.20	0.15	0.07	0.24
Grade 1									
TOCA-concentration problems	0.09	0.05	0.13	0.03	0.00	0.07	0.10	0.06	0.14
TOCA-disruptive behavior	0.11	0.07	0.15	0.04	−0.01	0.09	0.21	0.15	0.26
TOCA-prosocial behavior	0.51	0.42	0.60	0.44	0.33	0.55	0.14	0.09	0.19
TOCA-emotion dysregulation	0.09	0.05	0.13	0.02	−0.02	0.07	0.21	0.15	0.27
TOCA-internalization	0.17	0.11	0.24	0.05	−0.03	0.12	0.32	0.24	0.40
TOCA-family problems	0.21	0.14	0.28	0.18	0.10	0.25	0.09	0.06	0.13
TOCA-family involvement	0.21	0.14	0.29	0.03	−0.03	0.10	0.22	0.15	0.29
COMP-prosocial behavior	0.13	0.03	0.22	0.08	−0.03	0.18	0.13	0.04	0.22

(continued)

Table 8. (continued)

Outcome	Two-Level HLM			Three-Level HLM					
	ρ	Low	High	ρ_3	Low	High	ρ_2	Low	High
COMP-emotion regulation	0.14	0.04	0.24	0.08	−0.03	0.19	0.15	0.06	0.24
COMP-academic competence	0.06	0.00	0.12	0.05	−0.01	0.11	0.05	0.00	0.10
COMP-total scale	0.12	0.03	0.21	0.08	−0.02	0.18	0.11	0.04	0.19
Grade 2									
TOCA-concentration problems	0.05	0.03	0.08	0.02	−0.01	0.05	0.11	0.07	0.14
TOCA-disruptive behavior	0.12	0.07	0.17	0.05	0.00	0.09	0.16	0.11	0.21
TOCA-prosocial behavior	0.40	0.31	0.50	0.38	0.28	0.48	0.10	0.07	0.14
TOCA-emotion dysregulation	0.11	0.06	0.16	0.06	0.01	0.11	0.13	0.08	0.18
TOCA-internalization	0.19	0.13	0.26	0.10	0.02	0.19	0.29	0.21	0.37
TOCA-family problems	0.23	0.15	0.31	0.20	0.12	0.29	0.09	0.05	0.12
TOCA-family involvement	0.12	0.07	0.17	0.03	−0.02	0.09	0.22	0.15	0.28
COMP-prosocial behavior	0.05	0.00	0.10	0.03	−0.03	0.08	0.11	0.04	0.19
COMP-emotion regulation	0.06	0.00	0.11	0.04	−0.02	0.10	0.06	0.00	0.12
COMP-academic competence	0.07	0.01	0.13	0.06	0.00	0.13	0.03	−0.02	0.07
COMP-total scale	0.07	0.01	0.13	0.06	−0.01	0.12	0.06	0.00	0.12
Grade 3									
TOCA-concentration problems	0.07	0.03	0.10	0.03	−0.01	0.06	0.09	0.06	0.13
TOCA-disruptive behavior	0.10	0.06	0.14	0.06	0.01	0.10	0.13	0.09	0.17
TOCA-prosocial behavior	0.38	0.29	0.47	0.33	0.23	0.43	0.14	0.09	0.19
TOCA-emotion dysregulation	0.11	0.06	0.16	0.05	0.00	0.10	0.15	0.10	0.20
TOCA-internalization	0.21	0.13	0.30	0.05	−0.02	0.13	0.27	0.19	0.35
TOCA-family problems	0.19	0.12	0.26	0.15	0.07	0.22	0.10	0.06	0.15
TOCA-family involvement	0.14	0.09	.20	0.06	0.00	0.12	0.19	0.13	0.25

(continued)

Table 8. (continued)

Outcome	Two-Level HLM			Three-Level HLM					
	ρ	Low	High	ρ_3	Low	High	ρ_2	Low	High
Grade 4									
TOCA-concentration problems	0.07	0.04	0.11	0.03	−0.01	0.07	0.12	0.07	0.16
TOCA-disruptive behavior	0.11	0.07	0.16	0.07	0.02	0.12	0.12	0.08	0.17
TOCA-prosocial behavior	0.29	0.20	0.37	0.22	0.13	0.31	0.16	0.10	0.23
TOCA-emotion dysregulation	0.11	0.06	0.16	0.04	−0.02	0.09	0.17	0.10	0.23
TOCA-internalization	0.12	0.06	0.17	NA			NA		
TOCA-family problems	0.06	0.02	0.09	0.02	−0.01	0.05	0.09	0.05	0.13
TOCA-family involvement	0.15	0.08	0.21	NA			NA		
Grade 5									
TOCA-concentration problems	0.09	0.05	0.14	0.04	−0.01	0.08	0.15	0.10	0.20
TOCA-disruptive behavior	0.16	0.10	0.22	0.07	0.01	0.14	0.18	0.12	0.24
TOCA-prosocial behavior	0.34	0.24	0.45	0.29	0.17	0.41	0.18	0.11	0.24
TOCA-emotion dysregulation	0.12	0.07	0.18	0.01	−0.05	0.07	0.22	0.14	0.30
TOCA-internalization	0.15	0.08	0.21	0.01	−0.06	0.08	0.33	0.24	0.42
TOCA-family problems	0.03	0.01	0.05	NA			NA		
TOCA-family involvement	0.11	0.05	0.16	0.03	−0.02	0.09	0.22	0.15	0.29
M	0.15			0.09			0.16		
F		24.50***			29.21***			18.67***	
B		−0.01**			−0.02***			0.00	
r		−.11			−.16			.11	

Note. ICCs = intraclass correlations; HLM = hierarchical linear modeling; TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale; NA = not applicable. ρ is the school-level ICC for a two-level ICC, ρ_3 is the school-level ICC, and ρ_2 is the classrooms-level ICC for a three-level HLM; Low and high refer to the lower and higher bounds of the 95% confidence interval of ICCs, respectively. *M* refers to the mean. *F* refers to the *F*-statistic for the "outcome" variable and *b* refers to the regression coefficient of "grade" in the regression model. *r* refers to the Pearson correlation of ICCs and R^2 with grade across "outcome measures" (Hedges & Hedberg, 2007).
**$p < .01$. ***$p < .001$.

.22 to .49) than class-level ICCs (ρ_2 ranging from .10 to .18) from kindergarten to Grade 5, and the TOCA-C family problems measures had larger school-level ICCs (ρ_3 ranging from .15 to .20) than class-level ICCs (ρ_2 ranging from .09 to .13) from kindergarten to Grade 3. Most other TOCA-C and T-COMP measures had smaller school-level ICCs ($\rho_3 < .10$) than class-level ICCs ($\rho_2 > .10$) for students in kindergarten through Grade 5. The lower bounds of the 95% confidence intervals of ICCs for some measures were negative in three-level HLM, which indicated that the ICC estimates are not significantly different from 0 and there were not significant variance associated with those levels. The mean unconditional ICCs across outcome measures and grades at the school level (ρ) for two-level HLM is .15, and the mean ICCs at the school level (ρ_3) and class level (ρ_2) for a three-level HLM are .09 and .16, respectively. The Pearson correlations (r) between grades and ICCs are $-.11$ for two-level HLM, and $-.16$ and .11 for three-level HLM at the school level and class level, respectively. The F-statistics indicate that there were significant variances in ICCs associated with the outcome measures. The regression coefficients (b) of the predictor grade on school-level ICCs are $-.01$ for two-level HLM and $-.02$ for three-level HLM, and both are statistically significant at an α of .01.

Tables 9 and C3 report the proportion of variance explained by demographic covariates, pretest covariate, and demographic and pretest covariates at different levels for two- and three-level HLM (Level 1, R_1^2; Level 2, R_2^2; Level 3, R_3^2) by grade as well as the summary statistics for the pretest measured at the beginning of same school year and the pretest measured at the end of previous school year, respectively. Although R^2 at the same levels varied across outcome measures (all the F-statistics are significant at an α of .05), there were some general patterns, such that demographic covariates usually had smaller contribution in explaining variances than pretest. The mean R^2 across outcome measures and grades at the school and student levels are both .07 for two-level HLM, and the mean R^2 is .12 at the school level, .00 at class level, and .07 at the student level for a three-level HLM. In some instances, demographic covariates increased variances (negative R^2). However, when comparing among three levels, demographic covariates tended to explain a larger amount of variance at school level than at the class and student levels. Pretest covariates explained a large proportion of variance at all three levels (The mean R^2 across outcome measures and grades at the school and student levels are .52 and .42 for two-level HLM, and the mean R^2 is .51 at the school level, .46 at class level, and .40 at the student level for a three-level HLM), and there was no obvious difference based on whether the pretest was measured at the beginning of the

Table 9. R^2 for a Two- and Three-Level HLM by Grade: Pretest Measured at the Beginning (Fall) of Same School Year.

	Two-Level HLM						Three-Level HLM								
	Demographic Covariates Model		Pretest Covariate Model		Pretest and Demographic Covariate Model		Demographic Covariates Model			Pretest Covariate Model			Pretest and Demographic Covariate Model		
Outcome	R^2_2	R^2_1	R^2_2	R^2_1	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1
Kindergarten															
TOCA-concentration problems	0.00	0.09	0.51	0.49	0.46	0.51	0.05	−0.09	0.11	0.43	0.64	0.48	0.34	0.63	0.50
TOCA-disruptive behavior	0.12	0.05	0.72	0.48	0.72	0.49	0.18	0.00	0.06	0.93	0.55	0.47	0.92	0.55	0.48
TOCA-prosocial behavior	−0.01	0.04	0.90	0.35	0.89	0.37	−0.02	−0.01	0.05	0.91	0.68	0.28	0.90	0.68	0.30
TOCA-emotion dysregulation	−0.01	0.06	0.42	0.47	0.44	0.48	0.08	0.00	0.07	0.50	0.56	0.44	0.56	0.56	0.46
TOCA-internalization	−0.01	0.00	0.44	0.28	0.44	0.28	−0.01	0.00	0.00	0.62	0.49	0.21	0.63	0.49	0.21
TOCA-family problems	−0.11	0.04	0.37	0.16	0.27	0.17	−0.12	−0.03	0.04	0.34	0.36	0.13	0.23	0.34	0.16
TOCA-family involvement	0.10	0.09	0.52	0.35	0.48	0.37	0.24	−0.03	0.11	0.56	0.38	0.35	0.54	0.36	0.38
COMP-prosocial behavior	0.16	0.09	0.63	0.48	0.65	0.50	0.34	0.01	0.10	0.50	0.54	0.47	0.53	0.54	0.49
COMP-emotion regulation	−0.15	0.07	−0.06	0.36	−0.16	0.39	−0.64	0.01	0.09	−2.14	0.42	0.35	−2.64	0.42	0.38
COMP-academic competence	0.19	0.09	0.31	0.57	0.38	0.59	0.31	−0.22	0.10	0.40	0.23	0.59	0.50	0.21	0.61
COMP-total scale	0.15	0.10	0.39	0.56	0.39	0.58	0.28	−0.07	0.13	0.39	0.40	0.58	0.40	0.41	0.60
Grade 1															
TOCA-concentration problems	0.07	0.09	0.20	0.52	0.21	0.53	0.12	−0.02	0.09	−0.30	0.49	0.52	−0.27	0.47	0.53
TOCA-disruptive behavior	0.20	0.05	0.34	0.56	0.37	0.56	0.69	−0.03	0.06	0.81	0.61	0.54	0.93	0.60	0.54
TOCA-prosocial behavior	−0.03	0.03	0.82	0.39	0.82	0.40	−0.04	0.01	0.03	0.86	0.63	0.34	0.86	0.62	0.36
TOCA-emotion dysregulation	0.08	0.06	0.54	0.53	0.52	0.54	0.41	−0.01	0.07	0.64	0.71	0.50	0.52	0.71	0.51
TOCA-internalization	−0.01	0.01	0.34	0.36	0.33	0.36	0.03	−0.01	0.02	0.45	0.67	0.27	0.46	0.67	0.27
TOCA-family problems	−0.15	0.04	0.17	0.17	0.07	0.19	−0.19	0.01	0.04	0.16	0.27	0.16	0.04	0.27	0.18
TOCA-family involvement	0.14	0.08	0.24	0.39	0.30	0.40	0.47	0.03	0.09	0.92	0.26	0.39	1.00	0.28	0.41
COMP-prosocial behavior	0.15	0.09	0.72	0.45	0.83	0.47	0.13	0.11	0.09	0.95	0.24	0.47	1.00	0.34	0.49
COMP-emotion regulation	0.15	0.05	−0.07	0.34	0.04	0.35	0.22	0.06	0.05	−0.32	0.35	0.33	−0.16	0.37	0.34

(continued)

Table 9. (continued)

	Two-Level HLM						Three-Level HLM								
	Demographic Covariates Model		Pretest Covariate Model		Pretest and Demographic Covariate Model		Demographic Covariates Model			Pretest Covariate Model			Pretest and Demographic Covariate Model		
Outcome	R^2_2	R^2_1	R^2_2	R^2_1	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1	R^2_3	R^2_2	R^2_1
COMP-academic competence	0.46	0.08	0.89	0.63	1.00	0.64	0.67	−0.07	0.08	0.97	0.42	0.65	1.00	0.52	0.65
COMP-total scale	0.33	0.09	0.49	0.57	0.60	0.58	0.46	0.07	0.10	0.52	0.33	0.59	0.68	0.36	0.60
Grade 2															
TOCA-concentration problems	−0.11	0.11	0.01	0.49	−0.03	0.51	−0.36	0.03	0.12	−1.25	0.62	0.48	−1.33	0.60	0.50
TOCA-disruptive behavior	0.12	0.08	0.62	0.47	0.63	0.48	0.15	0.03	0.09	0.71	0.36	0.49	0.76	0.36	0.50
TOCA-prosocial behavior	−0.05	0.04	0.79	0.39	0.79	0.40	−0.06	0.03	0.04	0.84	0.54	0.37	0.83	0.55	0.38
TOCA-emotion dysregulation	−0.08	0.08	0.63	0.47	0.63	0.47	−0.15	0.00	0.09	0.69	0.38	0.48	0.69	0.38	0.49
TOCA-internalization	0.00	0.00	0.64	0.31	0.65	0.31	−0.02	−0.01	0.01	0.80	0.47	0.28	0.80	0.47	0.28
TOCA-family problems	−0.11	0.04	0.10	0.16	0.02	0.18	−0.14	−0.01	0.05	0.06	0.46	0.13	−0.04	0.45	0.16
TOCA-family involvement	0.09	0.09	0.53	0.38	0.53	0.39	0.18	0.04	0.10	0.28	0.58	0.35	0.35	0.57	0.37
COMP-prosocial behavior	−0.53	0.09	0.03	0.54	−0.09	0.54	−0.80	−0.04	0.10	−0.03	0.47	0.54	−0.24	0.48	0.55
COMP-emotion regulation	−0.36	0.08	0.01	0.38	−0.22	0.40	−0.46	−0.06	0.09	−0.04	0.27	0.39	−0.35	0.26	0.40
COMP-academic competence	−0.02	0.11	0.89	0.66	0.85	0.67	0.06	−0.59	0.13	1.00	−0.32	0.69	1.00	−0.37	0.70
COMP-total scale	−0.33	0.12	0.51	0.64	0.40	0.64	−0.37	−0.15	0.13	0.79	0.21	0.66	0.65	0.22	0.67
Grade 3															
TOCA-concentration problems	0.17	0.12	0.26	0.53	0.32	0.55	0.36	0.03	0.13	−0.33	0.77	0.52	−0.18	0.75	0.54
TOCA-disruptive behavior	0.22	0.08	0.69	0.55	0.68	0.55	0.42	−0.01	0.08	0.67	0.73	0.53	0.66	0.72	0.53
TOCA-prosocial behavior	−0.04	0.05	0.80	0.37	0.79	0.38	−0.05	−0.02	0.06	0.87	0.51	0.35	0.86	0.49	0.36
TOCA-emotion dysregulation	0.14	0.07	0.60	0.53	0.61	0.54	0.36	−0.02	0.08	0.52	0.72	0.51	0.56	0.71	0.52
TOCA-internalization	0.01	0.01	0.73	0.35	0.73	0.35	0.08	−0.01	0.01	0.89	0.60	0.29	0.88	0.60	0.29
TOCA-family problems	−0.17	0.04	0.43	0.16	0.34	0.17	−0.25	0.04	0.04	0.46	0.21	0.15	0.33	0.22	0.17
TOCA-family involvement	0.24	0.07	0.50	0.33	0.58	0.35	0.24	0.09	0.08	0.63	0.42	0.32	0.66	0.46	0.33

(continued)

Table 9. (continued)

	Two-Level HLM						Three-Level HLM								
	Demographic Covariates Model		Pretest Covariate Model		Pretest and Demographic Covariate Model		Demographic Covariates Model			Pretest Covariate Model			Pretest and Demographic Covariate Model		
Outcome	R_2^2	R_1^2	R_2^2	R_1^2	R_2^2	R_1^2	R_3^2	R_2^2	R_1^2	R_3^2	R_2^2	R_1^2	R_3^2	R_2^2	R_1^2
Grade 4															
TOCA-concentration problems	0.32	0.12	0.75	0.57	0.76	0.58	0.62	0.06	0.12	1.00	0.52	0.58	1.00	0.52	0.59
TOCA-disruptive behavior	0.43	0.09	0.74	0.51	0.80	0.51	0.68	0.04	0.09	0.96	0.51	0.51	1.00	0.54	0.51
TOCA-prosocial behavior	−0.02	0.05	0.81	0.28	0.83	0.29	−0.03	−0.01	0.06	0.98	0.26	0.29	0.99	0.26	0.31
TOCA-emotion dysregulation	0.21	0.08	0.71	0.50	0.72	0.50	0.65	−0.01	0.10	0.92	0.60	0.48	0.95	0.60	0.49
TOCA-internalization	0.05	0.01	0.73	0.27	0.74	0.27	NA	NA	NA	NA	NA	NA	NA	NA	NA
TOCA-family problems	−0.02	0.05	0.36	0.14	0.28	0.16	0.12	0.02	0.05	0.80	0.12	0.14	0.72	0.12	0.16
TOCA-family involvement	0.29	0.09	0.58	0.37	0.62	0.39	NA	NA	NA	NA	NA	NA	NA	NA	NA
Grade 5															
TOCA-concentration problems	0.38	0.10	0.81	0.60	0.83	0.60	0.65	0.13	0.10	0.97	0.68	0.59	0.99	0.70	0.59
TOCA-disruptive behavior	0.42	0.09	0.86	0.50	0.87	0.51	0.75	0.10	0.09	1.00	0.67	0.49	1.00	0.69	0.50
TOCA-prosocial behavior	−0.04	0.04	0.88	0.32	0.89	0.33	−0.06	0.03	0.05	0.98	0.48	0.29	0.99	0.48	0.30
TOCA-emotion dysregulation	0.21	0.10	0.84	0.47	0.82	0.48	−0.03	0.16	0.09	1.00	0.68	0.45	1.00	0.68	0.47
TOCA-internalization	0.06	0.01	0.76	0.33	0.75	0.33	−0.01	0.03	0.01	1.00	0.57	0.29	1.00	0.57	0.29
TOCA-family problems	0.22	0.03	0.13	0.14	0.21	0.15	NA	NA	NA	NA	NA	NA	NA	NA	NA
TOCA-family involvement	0.07	0.07	0.29	0.34	0.35	0.35	−0.27	0.07	0.08	−0.01	0.38	0.33	0.09	0.40	0.34
M	0.07	0.07	0.52	0.42	0.51	0.43	0.12	0.00	0.07	0.51	0.46	0.40	0.50	0.47	0.42
F	2.24*	41.74***	6.54***	71.32***	7.08***	80.36***	2.32*	4.21***	33.85***	4.10***	5.86***	94.85***	4.37***	4.92***	106.20***
b	0.03	0.01***	0.04*	−0.00	0.04*	0.00	0.01	0.01^	0.00	0.08^	−0.01	0.00	0.08^	−0.01	0.00
r	.24^	.02	.30*	−.13	.30*	−.14	.13	.34*	−.08	.29*	.07	−.12	.29*	.09	−.14

Note. HLM = hierarchical linear modeling; TOCA = Teacher Observation of Classroom Adaptation; COMP = Social Competence Scale; NA = not applicable. R_3^2, R_2^2, and R_1^2 are the proportions of variance explained by the covariates at school level (Level 3), teacher level (Level 2), and student level (Level 1), respectively. *M* refers to the mean. *F* refers to the *F*-statistic for the "outcome" variable and *b* refers to the regression coefficient of "grade" in the regression model. *r* refers to the Pearson correlation of ICCs and R^2 with grade across "outcome measures" (Hedges & Hedberg, 2007).
^$p < .10$. *$p < .05$. **$p < .01$, ***$p < .001$.

school year or whether the pretest was measured at the end of previous school year. The pretest and demographic covariates model, however, did not explain more variance than the pretest covariate model. The relationship between R^2 and grade varies across models and levels. The regression coefficients (b) of the predictor grade are .01 at the student level for the two-level demographic covariates model and .04 at the school level for the two-level pretest covariate model, and both are statistically significant at an α of .05.

Discussion

Interpreting the Intervention Effect Size Using the Empirical Benchmark

As Lipsey and colleagues (2012) suggest, when interpreting the effect size of an intervention, it is advisable to use multiple criteria including normative expectations for change and policy-relevant demographic performance gaps (also see Bloom et al., 2008; Hill et al., 2008). This study attempted to provide the normative expectations for change and existing demographic performance gaps on social and behavioral outcome measures to help policy makers define policy relevance and set realistic targets for interventions. In addition, policy makers, practitioners, and researchers can interpret the intervention effect sizes using these empirical benchmarks.

Because the TOCA-C measures did not illustrate consistent patterns on yearly changes, and most T-COMP measures illustrated significant changes from the beginning to the end of school years and same directional results, in particular, TOCA-C emotion dysregulation demonstrated opposed results with T-COMP emotion regulation on annual change; it appears that in studies over the course of a single year, it may be more advisable to interpret the effect size for the T-COMP scales than for the TOCA-C. For example, supposing that the effect size for an intervention on the T-COMP total scale is .20 at Grade 1, then based on the results in Table 3, we can interpret the effect sizes of this intervention as representing about 74% (=.20/.27) improvement over the annual gain for first graders. Similarly, using the results in Tables 4–7, we can interpret the effect size of an intervention regarding reducing demographic performance gaps among subgroups. For example, if an effect size of one intervention on the TOCA-C emotion dysregulation at Grade 2 is .15, it is equivalent to reducing the Black–White gap by 79% (=.15/.19; .19 is the Black–White gap at the end of Grade 2 on TOCA-C emotion dysregulation, see Table 4), reducing the

White–Hispanic gap by 75% (=.15/.20; .20 is the White–Hispanic gap at the end of Grade 2 on TOCA-C emotion dysregulation, see Table 5), reducing the boys–girls gap by 35% (=.15/.43; .43 is the boys–girls gap at the end of Grade 2 on TOCA-C emotion dysregulation, see Table 6), and reducing the FARMs ineligible–eligible gap by 125% (=.15/.12; .12 is the FARMs ineligible–eligible gap at the end of Grade 2 on TOCA-C emotion dysregulation, see Table 7). For further discussions about interpretations, see Lipsey et al. (2012).

Conducting Power Analyses Using the Parameter Reference Values

We also aimed to provide information that would inform sample size calculations when designing CRTs. Toward that end, we provide several examples of these design parameters in planning designs using a freeware that computes the minimum detectable effect size and minimum required sample size for multilevel CRTs using *PowerUp*! (Dong & Maynard, 2013), but most of the same calculations could be done using other software (e.g., CRT-Power: Borenstein, Hedges, & Rothstein, 2012; and Optimal Design: Raudenbush et al., 2011). For example, suppose that a researcher would like to design an experimental study to examine the effects of an intervention that aims to reduce children's emotion dysregulation at Grade 2. This is a school-level intervention, so the unit of random assignment is school. Suppose she expects that each school includes two classrooms, and each classroom includes 20 children. Hence, this is a three-level CRT design, where students (Level 1) are nested within classrooms (Level 2), and classrooms are nested with schools (Level 3) which is also the unit of the random assignment. The researcher would like to detect an effect size of .15 on the TOCA-C emotion dysregulation at Grade 2 with a statistical power of 80% if the intervention has effects. She would like to detect an effect size of .15 because she considers this effect size meaningful in prevention science, which is equivalent to reducing Black–White gap by 79% (as discussed earlier). The analytic model should match the study design, hence she plans to use a three-level HLM to estimate the intervention effect and two-sided test with an α of .05 for hypothesis testing. She wanted to know how many schools are needed for the study and used *PowerUp*! (Dong & Maynard, 2013) to conduct a power analysis. To calculate the sample size for schools, more assumptions that match the study design and analytic models are needed (e.g., ICCs at school and class levels). She can look for the school-level ICC (ρ_3) and class-level ICC (ρ_2) for the TOCA-C emotion dysregulation for the unconditional model in Tables 8 and C2, which both

provide the unconditional ICCs. It will not matter much if the two sets of ICCs differ little, but one can choose the bigger ICC to be conservative in power analysis. For three-level cluster random assignment designs, the Level 3 ICC is more critical on statistical power, so she can choose to use $\rho_3 = .06$ and $\rho_2 = .13$ from Table 8. Furthermore, she plans to use a balanced design, half of the schools are assigned to the treatment group ($p = .50$) and another half of the schools are assigned to the control group. She can then consider using covariates to improve statistical power. The proportions of variance explained by the demographic covariates (R^2) for the TOCA-C emotion dysregulation at Grade 2 in Tables 9 and C3 were very small, and some were negative. However, the R^2 for the pretest was very large. She decides to collect the pretest at the beginning of school year, and she can use $R_3^2 = .69$, $R_2^2 = .38$, and $R_1^2 = .48$ (Table 9) in the power analysis. After inputting the parameters assumed above in the spreadsheet "Model 3.2: Sample Size Calculator for 3-Level Cluster Random Assignment Designs (CRA3_3r)—Treatment at Level 3" in *PowerUp*!, she can click "RUN" to run the macro (more instruction is in Supplemental Materials A). The calculated sample for schools (kindergarten) is 99 (Demonstration 1, Table C4). Additional changes can be made within the *PowerUp*! spreadsheets based on different parameter assumptions (e.g., unbalanced study design, more classes in each school) and changing the reference design parameters (e.g., using different effect sizes or ICCs with justification). For example, the ICC estimates in Tables 8 and B2 have estimation errors, the upper bounds of the 95% confidence intervals of ρ_3 and ρ_2, that is .11 and .18 (Table 8), can be used to calculate the sample size in order to have a more conservatively large sample to detect the desired effect size with 80% power if the intervention has an effect.

We have demonstrated how to use the parameter reference values provided in this article to calculate the sample size for three-level simple CRT designs. The design parameters provided in this article also allow researchers to plan a two-level CRT where students are nested within schools. Using the same example above, suppose the classroom information is not available and the researcher would like to have 40 students per school, using similar power analysis approach above she will need 75 schools (Table C5). Zhu, Jacob, Bloom, and Xu (2012) conducted research about analyzing three-level CRTs by omitting the middle level and found that the treatment effect estimates and minimum detectable effect size (MDES) were similar between the three- and two-level models under some conditions. The example in this article (two classrooms per school), however, suggests that a two-level CRT has more power than a three-level CRT. The researchers may use

the design parameters provided in this article to conduct power analysis and select a design with the larger power. Furthermore, these design parameters can be used for another type of study design, three-level blocked CRT design (Level 2 is the unit of random assignment) and three-level blocked individual random assignment design (Level 1 is the unit of random assignment). See Appendix B for discussion of t.

Limitations

This is the first study to provide empirical benchmarks of meaningful effect sizes regarding the annual developmental change and policy-relevant demographic performance gaps and other design parameters (ICCs and R^2) on social and behavioral outcomes in prevention science; however, there are some limitations to consider. First, we used only data from four intervention studies in three states (Maryland, Missouri, and North Carolina). The design parameters estimated in this article can be used for the study designs using similar sample from these states and the sample from other similar states. These design parameters, however, may or may not be generalized to other populations. The effect size benchmarks and R^2 may be more generalizable than ICCs because effect sizes and R^2 are the characteristics of the outcome measures for individual students that may not vary a lot, while ICCs depends on the structures of schools and classrooms that may vary a lot across states.

Second, the sample size for some measures was not large enough to provide stable estimates of some design parameters, for example, the standard errors were large, the lower bounds of the 95% confidence intervals of ICCs were negative, some R^2 did not have consistent patterns across grades, and so on. For these measures, we suggest to use the mean design parameters as the proximate reference values for power analyses if there are no other better sources.

Third, the social and behavioral outcomes measures used in this article only included two teacher-rated scales, the TOCA-C and T-COMP. Duckworth and Yeager (2015) provided a detailed review of the limitations of teacher ratings in high-stakes accountability decisions (as well as a review of the limitations of performance and other measures of educational outcomes). In particular, they described the problem of reference bias, that is, the tendency for ratings to be influenced by the rater's frame of reference. A parallel literature in clinical psychology has emerged in recent years to better understand informant bias and discrepancies (de los Reyes & Kazdin, 2005). Rather than assuming discrepancies are rooted in error, modern

approaches have attempted to disentangle bias or error from actual contextual accuracy to better understand the meaning and value of these discrepancies. Additionally, even when teacher ratings are biased, abundant research since the 1960s has documented that these perspectives influence teacher interactions with students and ultimately how much they learn and grow (Rosenthal, 1994). Continued advances in understanding the strength and limitations of teacher ratings will guide how the findings in the present study are used by researchers.

The focus here on teacher ratings of student behaviors is justified because of their wide scale use and acceptance as indicators of youth socioemotional outcomes. Teacher-rated measures of student adjustment fill the corpus of social behavior literature and are accepted and used as primary outcome measures in rigorous school-based trials funded by rigorous federal funding agencies including the Institute for Education Sciences and the National Institute on Drug Abuse. One reason for their strong utility and predictive validity is that teachers benefit from broad normative comparisons of social and emotion behaviors based on their interactions with large number of same-aged children year to year. Still it is important to note that there are many other social and behavioral outcomes measures and methods worth investigating; the present study may serve as a model for future efforts in this regard. In summary, researchers should be cautioned when they use the results from this study to interpret the effect size of interventions or conducting power analysis. More studies using larger sample and additional social and behavioral outcomes measures to examine these parameters are needed.

Conclusions

This study provided reference values on (1) the empirical benchmark of meaningful effect sizes regarding the developmental annual change, and policy-relevant demographic performance gaps (race/ethnicity, gender, and SES); (2) ICCs for schools and classrooms; and (3) R^2 of demographic and pretest covariates at different levels (school, classrooms, and student) on social and behavioral outcomes. Although there are some limitations, and these effect size benchmarks and design parameters are not perfect estimates, this study contributes to prevention and educational science by providing empirical reference values of these parameters on social and behavioral outcomes to assist researchers to interpret the effect size of an intervention and conduct power analysis for designing CRTs examining social behavioral outcomes for children.

In summary, when interpreting the effect size of an intervention on social and behavioral outcomes, it is advisable to use multiple criteria including empirical benchmarks regarding annual change and policy-relevant demographic performance gaps as suggested by Lipsey et al. (2012). When planning a CRT on social and behavioral outcomes, a researcher should carefully select covariates. The demographic covariates may not increase power because of little or negative R^2, but usually the pretest has big R^2, which can increase power. The researcher should also be cautioned that the R^2 may be different at different levels and one should not use the R^2 estimated from one level for another level which may over- or underestimate the power. The researcher can use the design parameters provided in this study to conduct power analysis and compare power between two- and three-level CRTs to choose the one with bigger power. Furthermore, to be conservative, the researchers can use the upper bound of the 95% confidence intervals of ICCs for power analysis, which provides larger sample size estimate. In addition, although the design parameters (ICCs and R^2) were estimated using the data from multiple projects in which either the schools or classrooms were the units for random assignment to the treatment and control conditions, they can be used for designing studies in which students, classrooms, or schools are the units for random assignment because they are the common design parameters no matter which level the units for random assignment are at.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here for Project 1 was supported by the Institute of Educational Sciences (IES), U.S. Department of Education, through Grant R305A100342 to the University of Missouri, for Project 2 was supported by IES through Grant R305A150169, awarded to Dr. Murray, and for Projects 3 and 4 was supported by IES through Grants R324A07118 and R305A090307, and by NIMH through Grant R01 MH67948-1A1. The opinions expressed herein are those of the authors and not the funding agencies.

Supplemental Material

The online [appendices/data supplements/etc.] are available at http://journals.sagepub.com/doi/suppl/10.1177/0193841X16671283.

Note

1. Statistical significance is determined at an α of .05 through this article.

References

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods*, *18*, 155–176.

Berkhof, J., & Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics*, *29*, 201–218.

Bierman, K., Domitrovich, C., Nix, R., Gest, S., Welsch, J., Greenber, M., . . . Gill, S. (2008). *Child Development*, 1802–1817.

Bloom, H. S., Hill, C. J., Black, A. B., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328. doi:10.1080/19345740802400072

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.

Borenstein, M., Hedges, L. V., & Rothstein, H. (2012). *CRT-Power*. Teaneck, NJ: Biostat.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science*, *10*, 100–115.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*, 133–148.

Bradshaw, C. P., Pas, E. T., Goldweber, A., Rosenberg, M. S., & Leaf, P. J. (2012). Integrating school-wide positive behavioral interventions and supports with tier 2 coaching to student support teams: The PBIS plus model. *Advances in School Mental Health Promotion*, *5*, 177–193.

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2015). Examining variation in the impact of school-wide positive behavioral interventions and supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology*, *107*, 546–557.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study

(ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*, 489–507.

Conduct Problems Prevention Research Group. (2002). Evaluation of the first 3 years of the Fast Track prevention trial with children at high risk for adolescent conduct problems. *Journal of Abnormal Child Psychology*, *30*, 19–35.

Corrigan, A. (2003). *Teacher social competence scale, grade 8/year 9* (Fast Track Project Technical Report). Retrieved from the Fast Track Project website, http://www.fasttrackproject.org

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*, 24–67. doi:10.1080/19345747.2012.673143

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251.

Gifford-Smith, M. (2000). *Teacher social competence scale, grade 6/year 7 update* (Fast Track Project Technical Report). Durham, NC: Duke University.

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*, 346–380.

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

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489.

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlations in three- and four-level models. *Educational and Psychological Measurement*, *72*, 893–909. doi:10.1177/0013164412445193

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at http://ies.ed.gov/ncser/

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

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early

risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*, 599–641.

Johns Hopkins Prevention Intervention Research Center. (2006). *The first generation of JHU PIRC preventive intervention trials: Methods and measures*. Retrieved December 20, 2015, from http://www.jhsph.edu/prevention/Data/Cohort_1_and_2/Methods_and_Measures

Kelcey, B., & Phelps, G. (2013). Strategies for improving power in school randomized studies of professional development. *Evaluation Review*, *37*, 520–554.

Kellam, S. G., Ling, X., Merisca, R., Brown, C. H., & Ialongo, N. (1998). The effect of the level of aggression in the first grade classroom on the course and malleability of aggressive behavior into middle school. *Developmental Psychopathology*, *10*, 165–185.

Kellam, S. G., Mackenzie, A., Brown, C. H., Poduska, J. M., Wang, W., Petras, H., & Wilcox, H. C. (2011). The Good Behavior Game and the future of prevention and treatment. *Addiction Science & Clinical Practice*, 73–84.

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 265–288.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation-Checklist: Development and factor structure. *Measurement and Evaluation in Counseling and Development*, *42*, 15–30.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncser/

Murray, D. W., Rabiner, D., & Carrig, M. (2014, March). *Grade level effects of the incredible years teacher training program on emotion regulation and attention*. Paper presented at the 2014 meeting of the Society for Research on Educational Effectiveness, Washington, DC.

National Research Council & Institute of Medicine. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, DC: The National Academies Press.

Petras, H., Chilcoat, H. D., Leaf, P. J., Ialongo, N. S., & Kellam, S. G. (2004). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 88–96.

Racz, S. J., King, K. M., Wu, J., Witkiewitz, K., & McMahon, R. J., & The Conduct Problems Prevention Research Group. (2013). The predictive utility of a brief

kindergarten screening measure of child behavior problems. *Journal of Consulting and Clinical Psychology*, *81*, 588–599. doi:10.1037/a0032366

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). *Optimal design plus empirical evidence* (Version 3.0). Retrieved January 20, 2015, from http://www.wtgrantfoundation.org

Reinke, W. M., Herman, K. C., & Dong, N. (2014, March). *A group randomized evaluation of the incredible years teacher training program*. Paper presented at the Annual Meeting of Society for Research on Educational Effectiveness, Washington, DC.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, *3*, 176–179.

SAS Institute Inc. (2013). *SAS/STAT® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schaeffer, C. M., Petras, H., Ialongo, N., Masyn, K. E., Hubbard, S., Poduska, J., & Sheppard, K. A. (2006). Comparison of girls' and boys' aggressive-disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. *Journal of Consulting and Clinical Psychology*, *74*, 500–510.

Schochet, P. Z. (2008). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Spybrook, J., Hedges, L., & Borenstein, M. (2014). Understanding statistical power in cluster randomized trials: Challenges posed by differences in notation and terminology. *Journal of Research on Educational Effectiveness*, *4*, 384–406. doi: 10.1080/19345747.2013.848963

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*, 298–318. doi:10.3102/0162373709339524

Stormshak, E. A., Bierman, K. L., Bruschi, C., Dodge, K. A., & Coie, J. D. (1999). Conduct problems prevention research group. The relation between behavior problems and peer preference in different classroom contexts. *Child Development*, *70*, 169–182.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the national assessment of educational progress* (Statistical Analysis Report. NCES 2009-455). Washington, DC: National Center for Education Statistics.

Webster-Stratton, C. (1998). Preventing conduct problems in Head Start children: Strengthening parenting competencies. *Journal of Consulting and Clinical Psychology*, *66*, 715–730.

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585–602.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*, 490–519.

Zhu, P., Jacob, R., Bloom, H. S., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Education Evaluation and Policy Analysis*, *34*, 45–68.

Author Biographies

Nianbo Dong is an assistant professor in the Department of Educational, School, and Counseling Psychology at the University of Missouri. His research interests are statistical power analysis, causal inference, and the design and analysis of randomized experiments and quasi-experiments.

Wendy M. Reinke is a professor in the Department of Educational, School, and Counseling Psychology at the University of Missouri. Her research interests are in prevention of disruptive behaviors in children and youth, and dissemination and implementation of evidence-based interventions in schools.

Keith C. Herman is a professor in the Department of Educational, School, and Counseling Psychology at the University of Missouri. His research interests are in prevention of internalizing behaviors in children and youth and dissemination and implementation of evidence-based interventions in homes and schools.

Catherine P. Bradshaw is a professor and the associate dean for Research and Faculty Development at the Curry School of Education at the University of Virginia and the deputy director of the Johns Hopkins Center for the Prevention of Youth Violence and co-director of the Johns Hopkins Center for Prevention and Early Intervention. Her research focuses on the development and prevention of behavioral and mental health problems in schools.

Desiree W. Murray is a senior research scientist and associate director for Research at the Frank Porter Graham Child Development Institute and a research associate professor in the School of Education at the University of North Carolina at Chapel Hill. She is also an affiliate with the Sanford School of Public Policy at Duke University. Dr. Murray researches self-regulation development and evaluates school-based social-emotional interventions for students with disruptive behavior, including ADHD.