Title Page

Title:

Gender, Racial, and Socioeconomic Disparities on Social and Behavioral Skills for K-8 Students With and Without Interventions: An Integrative Data Analysis of Eight Cluster Randomized Trials

Authors and Affiliations:

Nianbo Dong (Corresponding Author)
University of North Carolina at Chapel Hill
dong.nianbo@gmail.com

Keith C. Herman
University of Missouri
hermanke@missouri.edu

Wendy M. Reinke
University of Missouri
reinkew@missouri.edu

Sandra Jo Wilson
Abt Associates.
sandra_wilson@abtassoc.com

Catherine P. Bradshaw
University of Virginia
cpb8g@virginia.edu

Suggested citation:

Dong, N., Herman, K. C., Reinke, W. M., Wilson, S. J., & Bradshaw, C. P. (2022). Gender, racial, and socioeconomic disparities on social and behavioral skills for K-8 students with and without interventions: An integrative data analysis of eight cluster randomized trials. *Prevention Science*. Advance online publication. doi: 10.1007/s11121-022-01425-w

Gender, Racial, and Socioeconomic Disparities on Social and Behavioral Skills for K-8 Students With and Without Interventions: An Integrative Data Analysis of Eight Cluster Randomized Trials

Abstract

Despite decades of concern about disparities in educational outcomes for low SES students and students of color, there has been limited rigorous study of programmatic approaches for reducing these disparities in elementary or middle schools. We conducted integrative data analysis (IDA) of the combined data from eight Institute of Education Sciences funded cluster randomized trials to address the research gaps on social and behavioral outcome disparities. The final analytic sample includes 90,880 students in varying grade levels from kindergarten to Grade 8 in 387 schools in 4 states (Maryland, Missouri, Virginia, and Texas). Two-level hierarchical linear modeling was used for multilevel moderation analysis. This study provided empirical evidence that there were significant gender, racial, and socioeconomic disparities on social and behavioral outcome measures for elementary and middle school students, the disparities significantly varied across schools, and the disparities could be reduced by interventions. We discussed our findings, implications for interpreting effect sizes of interventions using disparities as empirical benchmarks, and study limitations. We concluded with suggestions for future research.

Key words:

gender, racial, and socioeconomic disparities; integrative data analysis (IDA); multilevel moderation analysis; social and behavioral outcome

Gender, Racial, and Socioeconomic Disparities on Social and Behavioral Skills for K-8 Students With and Without Interventions: An Integrative Data Analysis of Eight Cluster Randomized Trials

Social and behavioral health is very important to students' wellbeing including their success in school and life. For example, social competence, emotional regulation, self-control, and positive affect are closely associated with student academic achievement, graduation, college attendance, employment, and earnings (e.g., Duncan & Magnuson, 2011; Lazowski & Hulleman, 2016; Segal, 2013; Yeager & Walton, 2011; Zins, et al., 2004). Students with higher social competence also have lower likelihood of future public assistance, criminal justice involvement, and mental health challenges in their young adulthood (Jones et al., 2015).

Student social and behavioral health also intersects with parent involvement in education to influence youth outcomes (Barnard, 2004; Sheridan et al., 2019). For instance, meta-analyses have revealed that parent engagement in education improves children's social-behavioral competence and mental health (Sheridan et al., 2019) and academic achievement and academic behaviors (Smith et al., 2019). One study found that teacher ratings of parent involvement during the elementary years significantly predicted student high school grades and completion (Barnard, 2004).

Prior research not only supports the broad influence of social and behavioral development and parental involvement on students' academic and life outcomes, but also demonstrates considerable gender, racial, and socioeconomic disparities on these social and behavioral outcomes. For example, Duncan and Magnuson (2011) revealed that the disparities between boys and girls, favoring girls, were over 0.40 standard deviation on attention and behavior problems, whereas Black-White disparities, favoring White students, were over 0.30 standard deviation in the first grade. The disparities were larger in the fifth grade. In addition, there is a similar pattern for the disparities on attention and behavior problems between students with the bottom and the top socioeconomic status (SES) favoring higher SES students. Reardon and Portilla (2016) also reported similar effect sizes of White-Black and income (the 90th percentile vs. 10th percentile) disparities on self-control, approaches to learning, and externalizing behavior in Kindergarten. Finally, patterns of family involvement in education also vary differentially based on student and family demographic characteristics (Stormont et al., 2013).

These social and behavioral outcome and parent involvement disparities have been receiving increasing attention, particularly as they relate to issues of equity and social justice (e.g., Ramirez et al, 2021). Policy makers, researchers, and practitioners are interested in interventions to improve student's social behavioral outcomes and reduce disparities, and much evaluation research has been conducted to test the effectiveness of the interventions to achieve these goals. For example, many large studies have reported statistically significant main and differential (moderator) effects of several universal interventions on student and parent involvement outcomes (Bradshaw et al., 2012; Durlak et al., 2011; Herman & Reinke, 2017; Herman et al., 2022; Reinke et al., 2018). However, the current status of disparities in social behavioral outcomes for students with regard to gender, race/ethnicity, and SES is not very clear. For instance, what are the magnitudes of the average disparities and the heterogeneity across schools? In addition, it is not well understood whether universal interventions can reduce disparities. For instance, pretest and the latent profiles based on pretest have been found to be significant moderators in several universal interventions (e.g., Bradshaw et al., 2015; Ialongo et al., 2019; Reinke et al., 2018); however, the demographic information such as gender, race, and free lunch status were not found to be significant moderators in some universal interventions (e.g., Domitrovich et al., 2016; Ialongo et al., 2019; Reinke et al., 2018; Herman et al., 2022a). One reason for non-significant findings on moderation (or reduction of disparities) could be insufficient statistical power due to limited sample size (Dong et al., 2018; Dong et al., 2021).

Furthermore, understanding these policy relevant disparities can help with interpreting the treatment effect sizes using empirical benchmarks (Bloom et al., 2008; Lipsey et al, 2012). For example, Dong et al. (2016) reported the gender, racial, and socioeconomic disparities on social and behavioral skills for elementary students using data from four cluster randomized trials (CRTs) and suggested to interpret the effect size of an intervention by comparing it with the disparities to indicate what percentage of the disparities can be reduced for the given treatment effect size.

To improve generalization and increase statistical power, integrative data analysis (IDA) or individual patient or participant meta-analysis (IPD) has been increasingly used in prevention science (e.g., Brunwasser & Gillham, 2018; Brown et al., 2013; Brown et al., 2018). IDA/IPD is an approach that pools the raw individual-level data together across multiple studies for synthesis analysis (Curran & Hussong 2009; Stewart & Parmar, 1993). IDA has also been found to have greater statistical power and be less biased than the analysis of aggregate data for moderation analysis (e.g., Dagne et al., 2016; Petkova et al. 2013).

The purpose of this study was to conduct IDA/IPD to address the research gaps on social and behavioral outcome and parent involvement disparities with regard to gender, race/ethnicity, and SES. Specifically, our research questions include: (1) What are the effect size of the disparities regarding gender, race/ethnicity (White vs. Black; White vs. Hispanic), and SES on social behavioral and parent involvement outcomes for elementary and middle school students? (2) Is there significant variation on the disparities across schools? (3) Did the interventions reduce disparities in the treatment group (moderated treatment effect)?

To explore the potential for subgroup effects based on SES, gender, and ethnicity, we used the data from these eight different CRTs, thereby providing us greater power to detect such effects which may be more challenging to discern in a single study (Brown et al., 2013; Brunwasser & Gillham, 2018). Specifically, IDA enables us to leverage multiple CRTs testing a range of classroom management and behavioral-focused preventive interventions which although not necessarily targeting such disparities, may have helped close some gaps in outcomes for students in these marginalized groups. We focused on teacher ratings of student adjustment and parent involvement because abundant research has found that teachers are excellent informants of student social, emotional, and behavioral problems (Reinke et al., 2008; Schaffer et al., 2003; Zima et al., 2005), and that teacher perceptions of parent involvement are potent predictors of student academic success (Barnard, 2004). We begin with a description of the sample of the eight studies, followed by the statistical methods used for data analysis in the methods section. We report the gender, racial, and SES disparities in the control and treatment groups, disparity difference between the treatment and control groups, disparity variation among schools, and the treatment effects on gender, racial, and SES subgroups in the results section. We then discuss the disparity results by comparing with existing literature, highlight the differential intervention effects in reducing disparities and the disparity heterogeneity, followed with application of disparities as empirical benchmark for interpreting effect sizes. Finally, we conclude with implications and suggestions for practice and future research aimed at reducing disparities in educational outcomes for students.

Methods

We used IDA to analyze the combined data from eight cluster randomized trials (CRTs) conducted in four states (Maryland, Missouri, Virginia, and Texas); all eight studies were funded by the US Department of Education, Institute of Education Sciences (IES). Multilevel modeling was used to account for the nested data structure to estimate the parameters of interest.

Sample

We used data from these eight IES-funded CRTs because all of them evaluated the effectiveness of school-based prevention interventions and used the same outcome measures. Most projects involved two-day teacher training, and followed with coaching for some projects. All eight projects included a primary outcome of teacher reports of students' behavior using the Teacher

Observation of Classroom Adaptation–Checklist (TOCA-C; Koth et al., 2009; Werthamer-Larsson et al., 1991). Three projects were conducted in Maryland, three in Missouri, one in Virginia, and one in Texas. The projects, samples, and outcome measures are summarized in Table 1 and described in greater detail below.

Project 1 focused on testing the effectiveness of a school-wide Positive Behavioral Interventions and Supports (PBIS) program using a CRT. The study randomly assigned 37 Maryland elementary schools in five school districts to either a treatment (21 schools) or control (16 schools) condition (Bradshaw et al., 2010, 2012b). 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 (Bradshaw & Kush, 2020; Koth et al., 2009; Werthamer-Larsson et al., 1991). The results from this project indicated that the treatment effect varied by the students' latent profiles based on pretest (Bradshaw et al., 2015).

Project 2 was a CRT, and was referred to as the PBISplus Trial (Bradshaw et al., 2012a). 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. 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.

Project 3 was a three-arm CRT, where 9 out of 27 Maryland public elementary schools were randomly assigned to each of three conditions: (1) standard setting (control), (2) the Good Behavior Game (GBG) (treatment 1), and (3) GBG+PATHS (Promoting Alternative Thinking Strategies) (treatment 2) (Domitrovich et al., 2016; Ialongo et al., 2019). Approximately 8,000 students in Grades K to 5 were included. The majority were Black and economically disadvantaged. The pretest and posttest of the TOCA-C were collected in fall and spring, respectively. The results from this project indicated that the treatment effect varied by the students' pretest but not by gender, race, or free lunch status (Domitrovich et al., 2016; Ialongo et al., 2019).

Project 4 was a CRT to test the effectiveness of the Incredible Years Teacher Classroom Management (IY TCM) program (Reinke, et al., 2018). The participants included 105 teachers and 1,818 students in kindergarten to Grade 3 from nine urban Missouri schools serving primarily Black students. 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 the TOCA-C. The results from this project indicated that the treatment effect varied by the students' pretest but not by gender, race, or free lunch status (Reinke, et al., 2018).

Project 5 was a study to develop an online coaching system for supporting elementary teachers in classroom management, referred as the web-based Classroom Check-Up (CCU) model (Reinke, et al., in press). In Phase 3, a pilot study with 39 teachers and 619 students in Missouri was used to evaluate the promise of the web-based CCU for enhancing teacher practice and student social and academic outcomes. Teachers were randomly assigned to receive the web-based CCU (n=20) versus standard practice (n=19). The pre- and posttest on the Teacher Observation of Classroom Adaptation Revised (TOCA-R) were collected.

Project 6 was a CRT to evaluate the efficacy of the CHAMPS (Conversation, Help, Activity, Movement, Participation, Success) classroom management program on the social behavioral and academic outcomes of middle school students in Missouri (Herman et al., 2022a; Herman et al., 2022b). A final sample included 102 teachers and 1450 students in Grades 6 to 8 in four cohorts. Teachers were randomly assigned to receive CHAMPS or to business-as-usual control group (51 interventions, and 51 control). Teachers rated student engagement, social skills, and classroom behaviors using the TOCA-R were collected in the fall of school years as the baseline pretest and in the spring school years as posttest. The results from this project indicated that the treatment effect did not vary by race (Herman, et al., 2022a).

Project 7 was a three-armed CRT to test the efficacy of a classroom behavior management strategy, the Good Behavior Game (GBG) (Poduska et al., 2012; Poduska & Kurki, 2014). One hundred and sixteen Grade 1 teachers in 32 schools in Texas across two cohorts were randomly assigned to three conditions: (1) GBG Basic, (2) GBG with Coach, and (3) business usual as control. The total number of students was 2,065. Teachers rated student concentration problem and disruptive behaviors using the TOCA-R were collected in the fall of school years as the baseline pretest and in the spring school years as posttest.

Project 8 was a CRT to test the efficacy of the combination of the Good Behavior Game (GBG) and My Teaching Partner (MTP) (Tolan et al, 2020); 156 teachers in 71 schools in Virginia were randomly assigned to receive the intervention (GBG + MTP training) or a business-as-usual control condition. Approximately 1,559 students from Kindergarten to Grade 3 participated in the study. The TOCA-C were collected for students in the fall of school years as the baseline pretest and in the spring school years as posttest.

We combined the data from all eight projects. Missing data is one of the biggest challenges in IDA because of the complexity of the data for IDA, e.g., study-level variability and changing relationships among variables over time, etc. (Siddique et al, 2018). Although some missing data procedures have been used in the IDA literature, there is no consensus whether these procedures can produce accurate results. For instance, Brown et al (2018) applied standard full-information maximum likelihood estimation (MLE) methods to handle missing in their IDA study of 19 trials, however, Brown et al (2018) is not a methodological study and it did not provide any diagnosis or effectiveness of MLE. Using the same data as Brown et al. (2018), Siddique et al. (2018) conducted a methodological study on applying multiple imputation to handle missing data, and they found that "our imputation model is not preserving all the relationships among the data" (p. S102) and "Even after reducing the scale of our application, we were still unable to produce accurate imputations of the missing values." (p. S95). In addition, the data in our study added another layer of complexity: students were nested within schools. It is a bigger challenge for missing data procedure to handle heterogeneity among schools. Given that there is no clear guidance on handling missing in IDA, and there were considerably small overall missing rates and small differential missing rates in most of our trials (Table 1), we decided to include all students that had the posttest score of the outcome and no missing data for at least one key predictor among race, gender, and status of free or reduced price lunch (FRPL). When there were more than two treatment conditions in one project, we created a binary treatment variable by assigning the control group (business as usual) as 0 and the other treatment groups as 1 because we aimed to test the average moderated treatment effect across all interventions. The final analytic sample included 90,880 students in varying grade levels from kindergarten to Grade 8 in 387 schools in 4 states (Maryland, Missouri, Virginia, and Texas). The student sample is relatively diverse: White (40.9%), Black (48.7%), Hispanic (5.7%), other race (4.7%); female (48.1%); eligible for FRPL (51.3%). Because some projects did not collect data on some outcome measures or some key predictors, the sample sizes for different analyses varied (Table 2 and Online Resource Tables S1-S3).

Variables

The outcome is measured by the *Teacher Observation of Classroom Adaptation-Checklist* (TOCA-

C; Bradshaw & Kush, 2020; Koth, Bradshaw, & Leaf, 2009), a nonclinical measure of children's behavior completed by teachers using a Likert scale (from 1 = *never* to 6= *almost always*). Seven subscales of the TOCA-C reported by Bradshaw and Kush (2020) include: (1) Concentration problems (inattentive and off-task behavior), (2) Aggressive/disruptive behavior (disobedient, disruptive, and aggressive behaviors), (3) Emotion regulation problems (or dysregulation: impulsivity, frustration, and anger), (4) Family involvement (caregivers' involvement in their child's school and parent's comfort in their relationship with the teacher), (5) Family problems (caregivers' degree of stability in home life and academic support of their children), (6) Internalization (extent to which the child feels nervous, fearful, sad, withdraw, and worries), and (7) Prosocial behaviors (positive social interactions). See more details on the items for each subscale in Bradshaw and Kush (2020). The psychometric properties of the TOCA have been well documented (Bradshaw & Kush, 2020). For example, these scales have a consistent factor structure over time (Koth et al., 2009), demonstrate strong internal consistency (the Cronbach's Alphas range from .89 to .96; Bradshaw et al., 2015) and high test–retest reliability (e.g., the correlations of test and retest over a 2-week period for aggressive behavior and concentration problems are .75 to .95, Werthamer-Larsson et al., 1991), relate to external criteria (e.g., child prosocial, attention problem, and aggression explained 50% of child-level variance of peer preference (equivalent to a multiple correlation of .71), Stormshak et al., 1999), are sensitive to relatively modest intervention effects (Ialongo et al., 1999), and have strong predictive validity (e.g., more aggressive boys were twice as likely as less aggressive boys to commit later violent acts, Petras et al., 2004). Racz et al. (2013) reported that higher TOCA kindergarten scores were associated with more behavior problems, lower social skills, and poorer school adjustment reported by multiple informants (teacher, parent, and child) at the end of elementary, middle, and high school. In addition, Bradshaw and Kush (2019) conducted a study to investigate differential item functioning (DIF) using the sample of 17,456 children in 45 schools (Project 2 in this study). They found that "all items on the Concentration, Aggressive/Disruptive Behavior, Emotion Regulation Problems, Family Problems, and Family Involvement subscales were shown to have little to no DIF for gender, race, and grade subgroups." (p.33) They concluded that results from the DIF analyses provide strong evidence of measurement invariance". (p.33) Bradshaw and Kush (2019) provide strong support for our study because our study focused on the gender, racial, and socioeconomic disparities on the TOCA measures.

The other relevant variables used in the analyses include student grade, race/ethnicity, gender, FRPL status (a proxy for SES), and treatment status.

Analytic Approach

We calculated the gender, racial, and socioeconomic disparities on the social and behavioral outcomes with and without receiving the interventions using two-level hierarchical linear models (HLM) to account for students nested within schools.

Level 1 (student):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \sum_{g=0}^{7} \beta_{(g+2)j}G_{gij} + e_{ij}, g = 0,,,7;\ e_{ij} \sim N(0, \sigma_{XG}^2) \qquad [1]$$

Level 2 (school):

$$\begin{aligned} &\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \sum_{p=2}^{8} \gamma_{0p}P_{pj} + \mu_{0j}, p = 2,,,8 \\ &\beta_{1j} = \gamma_{10} + \gamma_{11}T_j + \mu_{1j}, \binom{u_{0j}}{u_{1j}} \sim N\left[\binom{0}{0}, \begin{pmatrix} \tau_{00|TP}^2 & \\ \tau_{10|T} & \tau_{11|T}^2 \end{pmatrix}\right] \\ &\beta_{(g+2)j} = \gamma_{(g+2)0}, g = 0,,,7. \end{aligned} \qquad [2]$$

The combined model is:

$$Y_{ij} = \gamma_{00} + \gamma_{01}T_j + \sum_{p=2}^{8} \gamma_{0p}P_{pj} + \gamma_{10}X_{ij} + \gamma_{11}T_jX_{ij} + \sum_{g=0}^{7} \gamma_{(g+2)0}G_{gij} + \mu_{0j} + \mu_{1j}X_{ij} + e_{ij}. \qquad [3]$$

Y_{ij} is one of seven subscales of the TOCA-C measure for student *i* in school *j*. X_{ij} is the key binary predictor for female (female = 1, male = 0), White vs. Black (White = 1, Black = 0), White vs. Hispanic (White = 1, Hispanic = 0), or eligible for FRPL or not (eligible = 1, ineligible =0). G_{gij} is dummy variable for grade, ranging from kindergarten (0) to Grade 7 (Grade 8 as the reference grade). T_j indicates treatment status (treatment = 1, control = 0). P_{pj} is dummy variable for the projects, ranging from 2 to 8 (Project 1 as the reference project).

The parameter, γ_{01}, is the average treatment effect when X_{ij} = 0 (i.e., male, Black, Hispanic, or ineligible for FRPL), and γ_{11} indicates the treatment effect difference between female and male, Black and White, Hispanic and White, or ineligible for FRPL and eligible for FRPL, i.e., the moderated treatment effect, and γ_{01}+γ_{11} is the average treatment effect when X_{ij} = 1 (i.e., female, White, White, or eligible for FRPL). In addition, the parameter, γ_{10}, indicates the average disparity for students without receiving the intervention, γ_{11} also indicates the difference on disparity between students receiving and without receiving the intervention, and γ_{10}+γ_{11} indicates the average disparity for students receiving the intervention. We calculate the effect sizes of the disparities, the average treatment effect on subgroups, and moderated treatment effect by dividing the relevant parameter estimates by the pooled standard deviation of the outcome that is calculated from the two-level unconditional model. For example, $d = \gamma_{10}/\sqrt{\tau_{00}^2 + \sigma^2}$ is the effect size of the average disparity for students without receiving the intervention, where τ_{00}^2 and σ^2 are Levels 2 & 1 variances in the unconditional two-level model that does not include grade or predictor (*X*).

The parameter, $\tau_{11|T}^2$, indicates the variability of the disparity across schools after accounting for the treatment effect. We standardize the variability by using $\omega_x = \sqrt{\tau_{11|T}^2/(\tau_{00}^2 + \sigma^2)}$ to define the heterogeneity of disparity across schools (Dong et al, 2021; Dong, Kelcey, & Spybrook, 2020).

We use SAS PROC MIXED for the data analysis. We specify the EMPIRICAL option to use the sandwich estimator to adjust all standard errors and test statistics involving the fixed-effects parameters (SAS Institute Inc., 2018). The ESTIMATE statement with CL option is used to compute the relevant parameters and their 95% confidence intervals (CIs).

Results

The two-level HLM analysis results on seven TOCA subscales are summarized in Table 2 and Online Resource Tables S1-S3. Five parameters of interest include: the disparity in controls (γ_{10}), the disparity in treated ($\gamma_{10} + \gamma_{11}$), the treatment effect (γ_{01}) on the subgroup when the predictor *X* = 0, the treatment effect ($\gamma_{01} + \gamma_{11}$) on the subgroup when the predictor *X* = 1, and the moderated treatment effect (γ_{11}, i.e., the disparity difference between the treated and controls). We report their point estimates, standard errors (SE), *p*-values, and 95% confidence intervals (CIs). In addition, we report their effect sizes (ESs) and 95% CIs. The unconditional intraclass correlation coefficients (ICCs = $\tau_{00}^2/(\tau_{00}^2 + \sigma^2)$), and the sample sizes (students and schools) are also reported. Specifically, Table 2 and Online Resource Tables S1-S3 report these results regarding the status of FRPL (eligible vs. ineligible), gender (female vs. male), race (White vs. Black), and race (White vs. Hispanic), respectively. Furthermore, we summarize the results of the variance and heterogeneity of disparity across schools in Table 3. We report the point estimates, SE, *p*-values, and 95% CIs of the variance ($\tau_{11|T}^2$) of the disparity across schools conditional on the treatment status. In addition, we report the standardized heterogeneity coefficients (ω_x), their 95% CIs, and the total variance ($\tau_{00}^2 + \sigma^2$). An α of 0.05 was used to determine statistical significance in the results reported below.

Disparities

Figure 1 indicates the effect sizes and their 95% CIs of disparities between eligible and ineligible FRPL, and Online Resource Figures S1-S3 indicate the effect sizes and their 95% CIs of female-male, White-Black, and White-Hispanic disparities. Red dots indicate the disparities in the control group; Green triangles indicate the disparities in the treatment group; Blue diamonds indicate the differences in disparities between the treatment and control groups.

Eligible vs. Ineligible for Free or Reduced Price Lunch (FRPL)

There were significant disparities on all seven TOCA subscales between students eligible and ineligible FRPL in both control and treatment groups, and all favored ineligible for FRPL students (Table 2; Figure 1). For example, the effect sizes of disparities on family problems were 0.45 and 0.27 SD in the control and treatment groups, respectively, and there was a significant difference on the disparities between treatment and control groups (*d* = 0.18, *p* = 0.025). The absolute values of effect sizes of disparities on concentration problems, disruptive behavior, and emotion dysregulation ranged from 0.17 to 0.33 SD in control and treatment groups. The effect sizes of disparities on family involvement were very large, -0.55 and -0.50 SD in the control and treatment groups, respectively. The effect sizes of disparities on internalization were relatively smaller, ranging from 0.07 to 0.12 SD. The effect sizes of disparities on prosocial behavior were -0.22 and -0.18 SD in the control and treatment groups, respectively. Except for family involvement, there was no significant disparity difference on the other outcome measures between the treatment and control groups.

Females vs. Males

There were significant disparities on all seven TOCA subscales between females and males in both control and treatment groups, and all favored females, but there were no significant differences on the gender disparities between the treatment and control groups (Online Resource Table S1 & Figure S1). For example, the absolute values of effect sizes of gender disparities on concentration problems, disruptive behavior, emotion dysregulation, and prosocial behavior ranged from 0.30 to 0.43 standard deviation (SD) in control and treatment groups. The effect sizes of gender disparities on family involvement, family problems, and internalization were relatively smaller, ranging from 0.07 to 0.14 SD.

White vs. Black

There were significant disparities on all TOCA subscales except internalization between White and Black students in both control and treatment groups, and all favored White students, but there were no significant differences on the White-Black disparities between the treatment and control groups (Online Resource Table S2 & Figure S2). For example, the absolute values of effect sizes of White-Black disparities on concentration problems, disruptive behavior, emotion dysregulation, and prosocial behavior ranged from 0.21 to 0.36 SD in control and treatment groups. The effect sizes of White-Black disparities on family involvement were 0.38 and 0.45 SD in the control and treatment groups, respectively. The White-Black disparities on family problems were -0.13 and -0.19 SD in the control and treatment groups, respectively. The White-Black disparities on internalization were non-significant, -0.02 and 0.03 SD in the control and treatment groups, respectively.

White vs. Hispanic

The pattern of the disparities between White and Hispanic students was not the same as that between White and Black students. There were significant disparities on family involvement (0.36 and 0.40 SD) and internalization (0.15 and 0.18 SD) between White and Hispanic students in both control and treatment groups, and all favored White students; however, the significant disparities on disruptive behavior (0.09 and 0.13 SD) and emotion dysregulation (0.21 and 0.19 SD) between White and Hispanic students in both control and treatment groups all favored Hispanic students (Online Resource Table S3 & Figure S3). The White-Hispanic disparities on concentration problems were non-significant, 0.01 and 0.02 SD in the control and treatment groups, respectively. The White-Hispanic disparities on family problems was significant in the control group ($d = 0.09$, $p = 0.035$) but non-significant in the treatment group ($d = 0.06$, $p = 0.278$). In addition, there were no significant differences on the White-Hispanic disparities between the treatment and control groups.

Heterogeneity of Disparity

The variance ($\tau_{11|T}^2$) and heterogeneity (ω_x) of the disparities across schools were significant in most scenarios, and they varied across seven TOCA subscales and across four predictors (Table 3). For the female-male disparities, there were significant heterogeneity on concentration problems, disruptive behavior, emotion dysregulation, family problems, and prosocial behavior, ranging from 0.08 to 0.24. The heterogeneity on family involvement and internalization were not significant. For the eligible-ineligible for FRPL disparities, there were significant heterogeneity on all seven TOCA subscales, ranging from 0.10 to 0.24. We plotted the heterogeneity of disparity and 95% CI in Online Resource Figure S4. For the White-Black disparities, there was significant heterogeneity on all subscales except internalization, ranging from 0.07 to 0.24. For the White-Hispanic disparities, there was significant heterogeneity on concentration problems, disruptive behavior, emotion dysregulation, and family involvement, ranging from 0.13 to 0.16. The heterogeneity on family problems was not significant. In addition, there was no consistent estimate on the variance of the White-Hispanic disparity across schools on internalization or prosocial behavior, hence, we did not report them in Table 3.

Treatment Effects on Subgroups

Overall the average treatment effect sizes on subgroups tended to be smaller than the disparities and most of them were not statistically significant with some exceptions (Table 2 and Online Resource Tables S1-S3). In addition, the interventions generally did not have large or significant effects in reducing disparities that were indicated by the moderated treatment effect sizes. We summarize some significant findings below.

The effect sizes and 95% CIs of interventions for students eligible and ineligible for FRPL were plotted in Figure 2. The treatment effect on family problems was not significant for either ineligible ($d = 0.08$, $p = 0.132$) or eligible ($d = -0.10$, $p = 0.445$) for FRPL; however, the treatment effect size difference was significant and favored students eligible for FRPL ($d = -0.18$, $p = 0.025$). In addition, the treatment effect on prosocial behavior was significant for students eligible for FRPL ($d = 0.09$, $p = 0.025$). The other average treatment effects on students eligible or ineligible for FRPL or the moderated treatment effect were not significant.

The effect sizes and 95% CIs of interventions for females and males were plotted in Online Resource Figure S5. The treatment effect size on concentration problems for males was -0.07 SD (*p* = 0.017). The treatment effect sizes on internalization for males and females were -0.09 SD (*p* = 0.049) and -0.09 SD (*p* = 0.041), respectively. The treatment effect sizes on prosocial behavior for females was 0.08 SD (*p* = 0.017). The average treatment effects for males or females or the moderated treatment effects on other outcome measures were not significant.

The effect sizes and 95% CIs of interventions for White and Black students were plotted in Online Resource Figure S6. The treatment effect on internalization was significant for Black students (*d* = -0.10, *p* = 0.048). In addition, the treatment effect on prosocial behavior was significant for Black students (*d* = 0.10, *p* = 0.006). The other average treatment effects on White or Black students or the moderated treatment effect were not significant. In the analysis of the sample of White vs. Hispanic students, there was no significant treatment effect for White or Hispanic students on any of seven outcome measures, and there was no significant moderated treatment effect (Online Resource Table S3).

Discussion

The findings indicate that there were significant disparities on multiple social behavioral outcomes for students between females and males, White and Black, White and Hispanic, and ineligible and eligible for FRPL in both the control and treatment groups. For example, the effect sizes for gender disparities on concentration problems, disruptive behavior, emotion dysregulation, and prosocial behavior ranged from 0.30 to 0.43 SD, the effect sizes for SES disparities on family involvement and family problems ranged from 0.27 and 0.55 SD, the effect sizes for White-Black disparities on family involvement ranged from 0.38 to 0.45 SD, and the effect sizes for White-Hispanic disparities on family involvement ranged from 0.36 and 0.40 SD. All these large disparities favored students who were female, ineligible for FRPL, and White. However, a few disparities favored students of color; for example, the significant disparities on disruptive behavior (0.09 and 0.13 SD) and emotion dysregulation (0.21 and 0.19 SD) between White and Hispanic students in both control and treatment groups all favored Hispanic students. The effect sizes of the disparities for students between females and males and between White and Black were consistent with the literature (e.g., Duncan & Magnuson, 2011; Reardon & Portilla, 2016). The disparities for students ineligible and eligible for FRPL also echoed the income disparity results reported by Reardon and Portilla (2016).

In addition, the largest racial and socioeconomic disparities among all seven outcome measures appeared for family involvement. Specifically, the findings indicated that the Black, Hispanic, and eligible for FRPL students had much less family involvement than White and ineligible for FRPL students in both the control and treatment groups. These disparities are especially important because prior evidence suggests that teacher perceptions of parent involvement are potent predictors of student outcomes across development (Bakker et al., 2007; Barnard, 2004). For instance, teacher ratings of parent involvement in elementary predicted drop out and high school performance, more strongly than parent self-ratings of involvement (Barnard, 2004). Notably, other evidence suggests that teachers tend to rate the involvement of parents of color and of lower economic means more negatively than other parents and that these perceptions may be driven in part by teacher biases (Herman & Reinke, 2017; Stormont et al., 2013). Although we conceptualized parent involvement as a unidimensional variable in the present study based on the TOCA subscale that was used, other studies have found that teacher ratings of parent involvement may include perceptions of both quantity (e.g., how much or how often) and quality. Teacher judgments about parent involvement quality, including their sense of comfort and alignment with the parent ("I have a good relationship with the child's parent"), may be the aspects of involvement most susceptible to bias and interpretation (see Stormont et al., 2013; Herman & Reinke, 2017).

Many aspects of the school environment contribute to these disparities particularly when parents of different cultural backgrounds feel unwelcome and judged by educators (Stormshak et al., 2005). Thus, these disparities in family involvement across groups likely require intentional and strategic interventions to reduce educator biases and ensure that all parents perceive the school as welcoming, accessible, and open to their participation (Herman et al., 2017; Herman et al., 2021; Thompson et al., 2017).

There were significant treatment effects on some outcome measures for some subgroups. For example, the treatment effect sizes on prosocial behavior were significant for males ($d = 0.08$, $p = 0.0169$), students ineligible for FRPL ($d = 0.09$, $p = 0.0251$), and Black students ($d = 0.10$, $p = 0.0061$). This result is consistent with the findings on prosocial behavior on all sample in Reinke, Herman, and Dong (2018) ($d = 0.13$, $p = 0.038$). Furthermore, the interventions significantly reduced the disparities on some outcome measures between students eligible and ineligible FRPL in the treatment group. For example, there was a significant reduction ($d = -0.18$, $p = 0.025$) on the socioeconomic disparity on family problems in the treatment group ($d = 0.27$, $p < 0.0001$) compared to the control group ($d = 0.45$, $p < 0.0001$). It suggests that the interventions were more effective for the students eligible for FRPL. However, there were no differential treatment effects regarding gender or race. It suggests that the interventions may have the same effects across gender and race.

Finally, the significant heterogeneity of the disparities across schools indicates that the disparities varied a lot across schools. The disparities could be very large in some schools, and very small (close to 0) or reverse direction in other schools. For example, the average (mean) female-male disparity on concentration problems in the control groups was -0.43 SD and the heterogeneity coefficient of the disparity was 0.24. It suggests that the female-male disparity could be -0.67 SD for the schools with a disparity one SD above the mean, and the disparity could be -0.19 SD for the schools with a disparity one SD below the mean. The large average disparities are disconcerting, but the huge heterogeneity of disparities is even more disconcerting, as it suggests that the disparities in some schools were much larger. These heterogeneities point to the strong influence of social behavior contexts at each school that likely contribute to the disparities. Said another way, these school-specific disparities suggest the need for social contextual interventions to reduce disparities. Schools are not simply innocuous settings where disparities are observed, but rather they are dynamic environments that actively shape and create the observed disparities. For instance, starting at school entry, Black students receive significantly lower rates of positive interaction and attention in schools than White students, and these interaction patterns escalate teacher ratings of Black student disruptive behaviors and risk for punitive discipline (e.g., suspension) over time (Reinke et al., 2016; Bradshaw et al., 2010). On a promising note, heterogeneity of disparities imply that some schools provide environments that at least in part mitigate these disparities. Statistically, the school-level factors (e.g., school environments) may explain some variance in the heterogeneity of disparities; practically, some school-level interventions may reduce average disparities as well as heterogeneity of disparities. Possible school level leverage points that may influence these disparities include racial composition of teachers and students within buildings; proactive versus punitive discipline practices; school safety and climate; rates of student bullying and victimization; school size; principal leadership style; teacher-student and student-peer relationship quality; staff commitment to diversity and anti-racist policies and practices; and quality of instruction (Bradshaw et al., 2009; 2010).

Implications

The results of disparities reported in this paper can expand our understanding of the current status of gender, racial, and socioeconomic disparities on social and behavioral outcomes for K-8 students, and the impacts of interventions on improving social and behavioral outcomes for all students and reducing disparities. In addition, these disparities can serve as empirical benchmarks for interpreting the effect sizes of interventions found in other and future research (Bloom et al., 2008; Dong et al., 2016; Hill et al., 2008). Because the traditional, commonly used Cohen's small-medium-large distinctions for interpreting effect sizes are, at best, not very useful for decision-makers (a "small" effect size in one context may be a meaningful one in another). At worst, this terminology can be misleading (for instance, decision-makers may ignore "small" effects that might in fact be meaningful or bring about substantial cost savings). Bloom et al. (2008), Hill et al. (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. To apply the disparities for interpreting the effect sizes of intervention, Herman et al. (2020) for example, reported that the CHAMPS (Project 6 in this paper) showed an effect size of -0.14 standard deviation on concentration problems for middle school students, this effect size is equivalent to reducing the gender disparity by 32.5% (= $\frac{-0.14}{-0.43}$, where -0.43 was the gender disparity in the control group), the socioeconomic disparity by 41.9% (= $\frac{-0.14}{-0.33}$, where -0.33 was the disparity between students ineligible and eligible for FRPL), and the White-Black disparity by 53.9% (= $\frac{-0.14}{-0.26}$, where -0.26 was the White-Black disparity). An effect size of -0.14 was considered "small" according to Cohen's rule of thumb, but we can see that this effect size translates to non-trivial reduction in gender, socioeconomic, and racial disparities. In addition, translations like this may be more intuitive to consumers of research as well.

Limitation

Although we used the IDA to analyze the combined data from eight large CRTs in four states (Maryland, Missouri, Virginia, and Texas), and it has the advantages of increasing statistical power and better generalization (Brown et al., 2013; Brunwasser & Gillham, 2018), there are some limitations that need to be considered when interpreting the results. First, although the interventions in eight CRTs were all universal, school-based prevention interventions and shared some common features, the interventions were not the same. In the analysis, we arbitrarily created a binary variable to indicate the treatment status with 1 representing the treatment group. Hence, the average treatment effect (the coefficient of the treatment variable) should be interpreted as the average treatment effect across multiple interventions. This is why this analytic approach was also referred as the individual patient or participant meta-analysis (Stewart & Parmar, 1993). In addition, we controlled for the project and grade levels (dummy variables) in our analysis. Hence, all the analysis results represented the averages across eight projects from Kindergarten to Grade 8. It is possible that some projects were more effective and the interventions on some grades were more effective. One direction for future research is to explore the heterogeneity of disparities and treatment effect heterogeneity across projects and grade levels. It is also worth to note that the middle school context and early adolescence has unique characteristics relative to elementary school on the studied variables. Hence, another direction for future research is to examine the disparities in the elementary school and middle school separately and test if there is any difference.

Second, there were missing data on some variables in some projects, hence, the sample sizes varied across different analyses. For example, the analysis for the socioeconomic disparities did not included sample from Project 7 because the FRPL variable was not collected in Project 7 (Table 1). Similarly, the White samples were different in the analysis of White vs. Black from White vs. Hispanic, hence, the treatment effects for White students may be slightly different in two types of analyses. Third, we relied on teacher ratings of student behaviors; thus, the observed disparities may be a reflection of teacher perceptions and biases as well as objective differences between students. The rationale for focusing on teacher ratings is that they are the most common source of referrals for student social, emotional, and behavioral problems and for special education evaluations (Zima et al., 2005), and teacher ratings of students and parents accurately predict student social behavioral problems across development (Barnard, 2004; Reinke et al., 2008; Schaffer et al., 2003). Additionally, because they interact with large numbers of youth during their careers, teachers provide a valuable normative perspective on youth behaviors, and their ratings are viewed as the gold-standard assessment for a wide range of youth prosocial and disruptive behaviors (Lane et al., 2009). Finally, unlike studies that have revealed bias in teacher ratings of parent involvement, evidence is mixed regarding the presence of systematic bias in teacher ratings of students from different cultural backgrounds (Mason et al., 2014). On the one hand, teacher perceptions of disruptive behaviors are likely biased against students of color (Huang, 2018; Huang, 2020) and these biases contribute to higher rates of suspension and the school-to-prison pipeline for Black students (Eddy et al., 2020). On the other hand, one study found teacher ratings of concentration problems was more accurate for racial/ethnic minority students than for White students (Hosterman et al., 2008). Regardless, teacher perceptions of student social and behavior health are closely linked to student academic and life outcomes. Thus, reducing these disparities through bias reduction and social ecological interventions is a high priority if we are to create more equitable school conditions. Importantly, evidence from the present study indicated that in some cases interventions reduced disparities for youth who qualified for FRPL. Because all of the interventions in the present study involved teacher and/or whole school training in providing effective environments and none involved directly addressing teacher biases, these findings suggest that school social behavior interventions can lead to objective reductions in student and family problems, particularly favoring youth who qualify for FRPL.

Conclusion

Taken together, this study provided additional empirical evidence of significant gender, racial, and socioeconomic disparities in social and behavioral outcome measures of elementary and middle school students. Although the disparities significantly varied across schools, some disparities were reduced by the interventions tested. The large disparities and disparity heterogeneity across schools was particularly disconcerting. We call for more research on interventions to improve social and behavioral outcomes for all students, and in particular, the interventions for reducing disparities.

Compliance with Ethical Standards

Funding

This study was supported by a grant from the US Department of Education, Institute of Education Sciences (IES) [R305D190013]. The original data analyzed in this project also come from other IES funded projects [R305A090307, R324A07118, R305A080326, R305A100342, R305A130375, R305A130143, R305A090446, R305A190162/R305A130107]. The opinions expressed herein are those of the authors and not the funding agency. We would also like to thank the data providers of eight IES funded projects, with special thanks to Nicholas Ialongo, Philip Leaf, Jeanne Poduska, and Patrick Tolan.

Ethic approval

Not applicable

Conflicts of interest/Competing interests

Co-author Catherine Bradshaw is the editor of the journal *Prevention Science*, however, the peer-review process for this manuscript was managed by another associate editor not affiliated with this paper. The authors declare that they have no other conflicts of interest.

Consent to participate

Not applicable

Reference

Barnard, W. M. (2004). Parent involvement in elementary school and educational attainment. *Children and Youth Services Review, 26*, 39–62. doi: 10.1016/j.childyouth.2003.11.002

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*(4), 289-328. doi: 10.1080/19345740802400072

Bradshaw, C.P., Pas, E., Barrett, S., Bloom, J., Hershfeldt, P., Alexander, A., McKenna, M., and Leaf, P. (2012). A State-Wide Partnership to Promote Safe and Supportive Schools: The PBIS Maryland Initiative. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(4): 225–237. doi: 10.1007/s10488-011-0384-6.

Bradshaw, C.P., Mitchell, M.M., O'Brennan, L.M., & Leaf, P.J. (2010). Multilevel exploration of factors contributing to the overrepresentation of Black students in office disciplinary referrals. *Journal of Educational Psychology, 102*(2): 508–520. doi: 10.1037/a0018450

Bradshaw, C. P. & Kush, J. M. (2020). Teacher Observation of Classroom Adaptation-Checklist: Measuring Children's Social, Emotional, and Behavioral Functioning, *Children & Schools*, 42 (1), 29–40. https://doi.org/10.1093/cs/cdz022

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 PBISplus model, Advances in School Mental Health Promotion, 5:3, 177-193. doi: 10.1080/1754730X.2012.707429

Bradshaw, C. P., Sawyer, A. L., & O'Brennan, L. M. (2009). A social disorganization perspective on bullying-related attitudes and behaviors: The influence of school context. *American Journal of Community Psychology, 43(3*), 204-220. doi: 10.1007/s10464-009-9240-1.

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2012). Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics, 130*, e1136–e1145. http://dx.doi.org/10.1542/peds.2012-0243

Bradshaw, C.P., Waasdorp, T.E., and 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*(2): 546–557. doi: 10.1037/a0037630

Brunwasser, S. M., Gillham, J. E. (2018). Identifying moderators of response to the Penn Resiliency Program: A synthesis study. *Prevention Science. 19*, 38-48. doi: 10.1007/s11121-015-0627-y.

Brown, C. H., Brincks, A., Huang, S., Perrino, T., Cruden, G., Pantin, H., Howe, G., Young, J. F., Beardslee, W., Montag, S., & Sandler, I. (2018). Two-year impact of prevention programs on adolescent depression: An integrative data analysis approach. *Prevention Science*, *19*(Suppl 1), 74–94. doi: 10.1007/s11121-016-0737-1

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burkhart, G. et al. (2013). Methods for Synthesizing Findings on Moderation Effects Across Multiple Randomized Trials. *Prevention Science, 14* (2), 144-156. doi: 10.1007/s11121-011-0207-8

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*, 81–100. doi: 10.1037/a0015914.

Dagne, G.A., Brown, C.H., Howe, G., Kellam, S.G., & Liu, L. (2016). Testing moderation in network meta-analysis with individual participant data. *Statistics in Medicine, 34*, 2485–2502. doi: 10.1002/sim.6883.

Domitrovich, C., Bradshaw, C. P., Berg, J., Pas, E. T., Becker, K., Musci, R., Embry, D. D. and Ialongo, N. (2016). How Do School-Based Prevention Programs Impact Teachers? Findings from a Randomized Trial of an Integrated Classroom Management and Social-Emotional Program. *Prevention Science, 17*(3): 325–337. doi: 10.1007/s11121-015-0618-z.

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses of moderator effects in three-level cluster randomized trials. *The Journal of Experimental Education, 86* (3), 489-514. doi: 10.1080/00220973.2017.1315714

Dong, N., Kelcey, B., & Spybrook, J. (2020). Design considerations in multisite randomized trials to probe moderated treatment effects. *Journal of Educational and Behavioral Statistics*. Advance online publication. doi: 10.3102/1076998620961492

Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for panning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review, 40*(4), 334-377. doi: 10.1177/0193841X16671283

Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (2021). Power analyses for moderator effects with (non)random slopes in cluster randomized trials. *Methodology, 17* (2), 92-110. doi: https://doi.org/10.5964/meth.4003

Duncan, G. J. & Magnuson, K. (2011). The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems, in Greg J. Duncan and Richard J. Murnane (eds.), *Whither Opportunity: Rising Inequality, Schools, and Children's Life Chances*, New York: Russell Sage, 2011, pp. 47-69.

Durlak, J., Weissberg, R., Dymnicki, A., Taylor, R., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*, 405–432. doi: 10.1111/j.1467-8624.2010.01564.x

Eddy, C. L., Huang, F. L., Cohen, D. R., Baker, K. M., Edwards, K. D., Herman, K. C., & Reinke, W. M. (2020). Does teacher emotional exhaustion and efficacy predict student discipline sanctions? *School Psychology Review*, 49(3), 239-255. doi: 10.1080/2372966X.2020.1733340

Herman, K. C., Dong, N., Reinke, W. M., & Bradshaw, C. P. (2022). Accounting for traumatic historical events in randomized controlled trials. *School Psychology Review*. Advance online publication. doi: 10.1080/2372966X.2021.2024768

Herman, K. C., & Reinke, W. M. (2017). Improving teacher perceptions of parent involvement patterns: Findings from a group randomized trial. *School Psychology Quarterly, 32,* 89-102. doi: 10.1037/spq0000169.

Herman, K. C., Reinke, W. M., Dong, N., & Bradshaw, C. (2022). Can effective classroom behavior management increase student achievement in middle school? Findings from a group randomized trial. *Journal of Educational Psychology*, *114*(1), 144–160.. doi: 10.1037/edu0000641

Herman, K. C., Reinke, W. M., & Frey, A. (2021). Motivational interviewing in schools: Strategies for engaging parents, teachers, and students (2nd edition). Springer Publishing Company, LLC.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2* (3), 172–177. doi: 10.1111/j.1750-8606.2008.00061.x

Hosterman, S. J., DuPaul, G. J., & Jitendra, A. K. (2008). Teacher ratings of ADHD symptoms in ethnic minority students: Bias or behavioral difference? *School Psychology Quarterly, 23(3)*, 418-435. doi: 10.1037/a0012668

Huang, F. L. (2018). Do Black students misbehave more? Investigating the differential involvement hypothesis and out-of-school suspensions. *The Journal of Educational Research, 111(3)*, 284-294. doi: 10.1080/00220671.2016.1253538

Huang, F. L. (2020). Prior problem behaviors do not account for the racial suspension gap. *Educational Researcher, 49(7),* 493-502. doi: 10.3102/0013189X20932474

Ialongo, N. S., Domitrovich, C., Embry, D., Greenberg, M., Lawson, A., Becker, K. D., Bradshaw, C. A (2019). Randomized controlled trial of the combination of two school-based universal preventive interventions. *Developmental Psychology. 55*(6):1313-1325. doi: 10.1037/dev0000715.

Jones, D. E., Greenberg, M., Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 105*(11), 2283–2290. doi: 10.2105/AJPH.2015.302630

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*(1), 15-30. doi: 10.1177/0748175609333560

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 17(2),* 93-105. doi: 10.1177/1063426609341069

Lazowski, R. A. & Hulleman, C. S. (2016). Motivation Interventions in Education: A Meta-Analytic Review. *Review of Educational Research, 86* (2), 602-640. doi: 10.3102/0034654315617832

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. (NCSER 2013-3000).

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools, 51(10),* 1017-1030. doi: 10.1002/pits.21800

McCoach, D., Goldstein, J., Behuniak, P., Reis, S. M., Black, A. C., Sullivan, E. E., & Rambo, K. (2010). Examining the unexpected: Outlier analyses of factors affecting student achievement. *Journal of Advanced Academics, 21,* 426–468. doi: 10.1177/1932202X1002100304

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. doi: 10.1097/00004583-200401000-00018

Petkova, E., Tarpey, T., Huang, L., & Deng, L. (2013). Interpreting metaregression: Application to recent controversies in antidepressants' efficacy. Statistics in Medicine, 32, 2875–2892. doi:10.1002/sim.5766.

Poduska, J. M., Gomez, M., Capo, Z., & Holmes, V. (2012). Developing a Collaboration With the Houston Independent School District: Testing the Generalizability of a Partnership Model. *Administration and Policy in Mental Health and Mental Health Services Research, 39*(4): 258–267. doi: 10.1007/s10488-011-0383-7

Poduska, J. M., & Kurki, A. (2014). Guided by Theory, Informed by Practice: Training and Support for the Good Behavior Game, a Classroom-Based Behavior Management Strategy. *Journal of Emotional and Behavioral Disorders, 22*(2): 83–94. doi: 10.1177/1063426614522692

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

Ramirez, T., Brush, K., Raisch N., Bailey, R. & Jones, S.M. (2021). Equity in Social Emotional Learning Programs: A Content Analysis of Equitable Practices in PreK-5 SEL Programs. *Frontiers in Education, 6*:679467. doi: 10.3389/feduc.2021.679467

Reardon, S.F., Portilla, X.A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, 2 (3). doi: 10.1177/2332858416657343

Reinke, W.M., Herman, K.C., & Copeland, C. (in press). Student engagement: The importance of the classroom context. In A. Reschly & S. Christenson (Eds.). *The Handbook of Research on Student Engagement* (second edition). New York: Springer.

Reinke, W. M., Herman, K. C., & Dong, N. (2018). The Incredible Years Teacher Classroom Management program: Outcomes from a group randomized trial. *Prevention Science, 19* (8), 1043–1054. doi: 10.1007/s11121-018-0932-3

Reinke, W.M., Herman, K. C., & Newcomer, L. (2016). The Brief Student-Teacher Interaction Observation: Using dynamic indicators of behaviors in the classroom to predict outcomes and inform practice. *Assessment for Effective Intervention, 42*, 32-42.

Reinke, W. M., Herman, K. C., Petri, H., & Ialongo, N. S. (2008). Empirically-derived subtypes of child academic and behavior problems: Co-occurrence and distal outcomes. *Journal of Abnormal Child Psychology, 36,* 759-770. doi: 10.1177/1534508416641605

SAS Institute Inc. (2018). SAS/STAT® 15.1 User's Guide. Cary, NC: SAS Institute Inc.

Schaeffer, C. M., Petras, H., Ialongo, N., Poduska, J., & Kellam, S. (2003). Modeling growth in boys aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology, 39,* 1020 –1035. doi: 10.1037/0012-1649.39.6.1020

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association, 11* (4), 743-779. doi: https://doi.org/10.1111/jeea.12025

Sheridan, S. M., Smith, T. E., Kim, E. M., Beretvas, S. N., & Park, S. (2019). A meta-analysis of family-school interventions and children's social-emotional functioning: Child and community influences and components of efficacy. *Review of Educational Research, 89,* 296-332. doi: 10.3102/0034654318825437

Siddique, J., de Chavez, P.J., Howe, G. et al. (2018). Limitations in Using Multiple Imputation to Harmonize Individual Participant Data for Meta-Analysis. *Prevention Science, 19*(Suppl 1), 95–108. doi: 10.1007/s11121-017-0760-x

Smith, T. E., & Sheridan, S. M. (2019). The effects of teacher training on teachers' family engagement practices, attitudes, and knowledge: A meta-analysis. *Journal of Educational and Psychological Consultation, 29,* 128-157. doi: 10.1080/10474412.2018.1460725

Stewart, L. A., & Parmar, M. K. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *Lancet* (London, England), 341, 418–422. doi: 10.1016/0140-6736(93)93004-k

Stormont, M., Herman, K. C., Reinke, W. M., David, K., & Goel, N. (2013). Latent profile analysis of teacher perceptions of parent contact and comfort. *School Psychology Quarterly, 28,* 195–209. doi: 10.1037/spq0000004

Stormshak, E. A., Dishion, T. J., Light, J., & Yasui, M. (2005). Implementing family-centered interventions within the public middle school: Linking service delivery to change in student problem behavior. *Journal of Abnormal Child Psychology, 33,* 723–733. doi: 10.1007/s10802-005-7650-6

Stormshak, E. A., Bierman, K. L., Bruschi, C., Dodge, K. A., & Coie, J. D. (1999). The relation between behavior problems and peer preference in different classroom contexts. Conduct problems prevention research group. *Child Development, 70*, 169–182. doi: 10.1111/1467-8624.00013

Thompson, A., Herman, K. C., Stormont, M., Reinke, W. M., & Webster-Stratton, C. (2017). Impact of Incredible Years on teacher perceptions of parent involvement: A latent transition analysis. *Journal of School Psychology, 62*, 51-65. doi: 10.1016/j.jsp.2017.03.003

Tolan, P., Elreda, L. M., Bradshaw, C. P., Downer, J. T., & Ialongo, N. (2020). Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: The moderating role of distress among new teachers on student outcomes. *Journal of School Psychology, 78*, 75-95. doi: 10.1016/j.jsp.2019.12.002

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. doi: 10.1007/BF00937993

Yeager, D. S., & Walton, G. M. (2011). Social-Psychological Interventions in Education: They're Not Magic. *Review of Educational Research, 81*, 267-301. doi: 10.3102/0034654311405999

Zima, B. T., Hurlburt, M. S., Knapp, P., et al. (2005). Quality of publicly funded outpatient specialty mental health care for common childhood psychiatric disorders in California. *Journal of American Academy of Child and Adolescent Psychiatry, 44(2),* 130–144. doi: 10.1097/00004583-200502000-00005

Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (Eds.). (2004). *Building academic success on social and emotional learning: What does the research say?* Teachers College Press.

Table 1

Descriptive Statistics of the Analytic Sample

Project	White (%)	Black (%)	Hispanic (%)	Other Race (%)	Female (%)	Free or reduced price lunch (%)	Number of Students	Number of Schools	Grade Level	State of Sample	Student Attrition/ Outcome Missing Rates	TOCA Subscales	IES Award Number
1	50.4	40.9	3.4	5.4	47.5	48.7	46,436	148	K-5	Maryland	8.4% (Control), 8.7% (Treated)	Concentration, Disruptive, Prosocial behavior	R305A090307
2	35.3	52.5	7.8	4.4	48.5	44.0	32,209	87	K-5	Maryland	7-11% across 3 years (Overall)	All seven	R324A070118
3	4.4	90.6	4.1	0.9	49.3	92.0	5,142	27	K-5	Maryland	6.1% (Control), 4.4-5.6% (Treated)	Concentration, Disruptive behavior	R305A080326
4	22.1	75.2	2.1	0.6	48.6	60.5	1,612	9	K-3	Missouri	7.6% (Control), 7.4% (Treated)	All seven	R305A100342
5	79.9	6.5	4.9	8.7	46.8	93.7	619	4	K-5	Missouri	4.9% (Overall)	All seven	R305A130375
6	19.2	76.6	2.2	2.0	50.6	67.8	1,244	9	6-8	Missouri	13.0% (Control), 15.6% (Treated)	All seven	R305A130143
7	43.5	26.5	28.9	4.9	50.2	NA	2,059	32	1	Texas	NA	Concentration, Disruptive behavior	R305A090446
8	11.0	62.2	17.3	9.5	50.7	77.2	1,559	71	K-3	Virginia	NA	All but Family problems	R305A130107
Total Sample	40.9	48.7	5.7	4.7	48.1	51.3	90,880	387	K-8				

Table 2: Two-level HLM Result Summary Regarding Free/Reduced Price Lunch (Eligible vs. Ineligible)

Outcome	Parameters of Interest	Estimate	SE	*P* value	95% CI		Effect Size (ES)	ES 95% CI		Unconditional ICC	Number of Students	Number of Schools
Concentration problems	Disparity in Controls (γ_{10})	0.41	0.03	<0.0001	0.36	0.46	0.33	0.29	0.37	0.04	66,854	331
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.35	0.02	<0.0001	0.30	0.40	0.29	0.25	0.33			
	Treatment Effect on Ineligible (γ_{01})	0.02	0.04	0.6775	-0.06	0.09	0.01	-0.05	0.07			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	-0.04	0.03	0.1749	-0.10	0.02	-0.03	-0.08	0.02			
	Moderated Treatment Effect (γ_{11})	-0.06	0.04	0.1063	-0.13	0.01	-0.05	-0.10	0.01			
Aggressive/ Disruptive behavior	Disparity in Controls (γ_{10})	0.20	0.01	<0.0001	0.18	0.23	0.24	0.21	0.27	0.06	66,839	331
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.17	0.01	<0.0001	0.14	0.20	0.20	0.17	0.23			
	Treatment Effect on Ineligible (γ_{01})	0.01	0.03	0.6204	-0.04	0.07	0.02	-0.05	0.08			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	-0.02	0.03	0.3863	-0.07	0.03	-0.03	-0.08	0.03			
	Moderated Treatment Effect (γ_{11})	-0.04	0.02	0.0826	-0.07	0.00	-0.04	-0.09	0.01			
Emotion Dysregulation	Disparity in Controls (γ_{10})	0.22	0.02	<0.0001	0.17	0.26	0.20	0.16	0.25	0.05	33,808	169
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.18	0.02	<0.0001	0.13	0.22	0.17	0.12	0.21			
	Treatment Effect on Ineligible (γ_{01})	-0.02	0.04	0.5893	-0.09	0.05	-0.02	-0.09	0.05			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	-0.06	0.05	0.1937	-0.15	0.03	-0.06	-0.14	0.03			
	Moderated Treatment Effect (γ_{11})	-0.04	0.03	0.2144	-0.10	0.02	-0.04	-0.10	0.02			
Family involvement	Disparity in Controls (γ_{10})	-0.76	0.04	<0.0001	-0.85	-0.67	-0.55	-0.61	-0.48	0.10	33,565	169
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	-0.69	0.05	<0.0001	-0.80	-0.59	-0.50	-0.58	-0.43			
	Treatment Effect on Ineligible (γ_{01})	0.00	0.07	0.9464	-0.13	0.14	0.00	-0.09	0.10			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	0.07	0.06	0.2256	-0.04	0.19	0.05	-0.03	0.13			
	Moderated Treatment Effect (γ_{11})	0.07	0.07	0.3374	-0.07	0.20	0.05	-0.05	0.15			
Family problems	Disparity in Controls (γ_{10})	0.31	0.03	<0.0001	0.25	0.38	0.45	0.35	0.54	0.04	32,629	102
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.19	0.04	<0.0001	0.12	0.26	0.27	0.17	0.37			
	Treatment Effect on Ineligible (γ_{01})	0.05	0.04	0.1324	-0.02	0.12	0.08	-0.02	0.18			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	-0.07	0.06	0.2516	-0.19	0.05	-0.10	-0.28	0.07			
	Moderated Treatment Effect (γ_{11})	-0.12	0.06	0.0251	-0.23	-0.02	-0.18	-0.34	-0.02			
Internalization	Disparity in Controls (γ_{10})	0.10	0.02	<0.0001	0.06	0.13	0.12	0.08	0.16	0.05	33,808	169
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.05	0.02	0.0006	0.02	0.09	0.07	0.03	0.11			
	Treatment Effect on Ineligible (γ_{01})	-0.03	0.03	0.4454	-0.09	0.04	-0.03	-0.11	0.05			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	-0.07	0.04	0.0944	-0.15	0.01	-0.08	-0.18	0.01			
	Moderated Treatment Effect (γ_{11})	-0.04	0.02	0.0705	-0.09	0.00	-0.05	-0.11	0.00			
Prosocial behaviors	Disparity in Controls (γ_{10})	-0.18	0.02	<0.0001	-0.22	-0.15	-0.22	-0.27	-0.18	0.08	62,008	304
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	-0.14	0.02	<0.0001	-0.18	-0.11	-0.18	-0.22	-0.13			
	Treatment Effect on Ineligible (γ_{01})	0.04	0.04	0.3986	-0.05	0.12	0.05	-0.06	0.15			
	Treatment Effect on Eligible ($\gamma_{01}+\gamma_{11}$)	0.08	0.03	0.0251	0.01	0.14	0.09	0.01	0.17			
	Moderated Treatment Effect (γ_{11})	0.04	0.03	0.1581	-0.02	0.09	0.05	-0.02	0.11			

Table 3

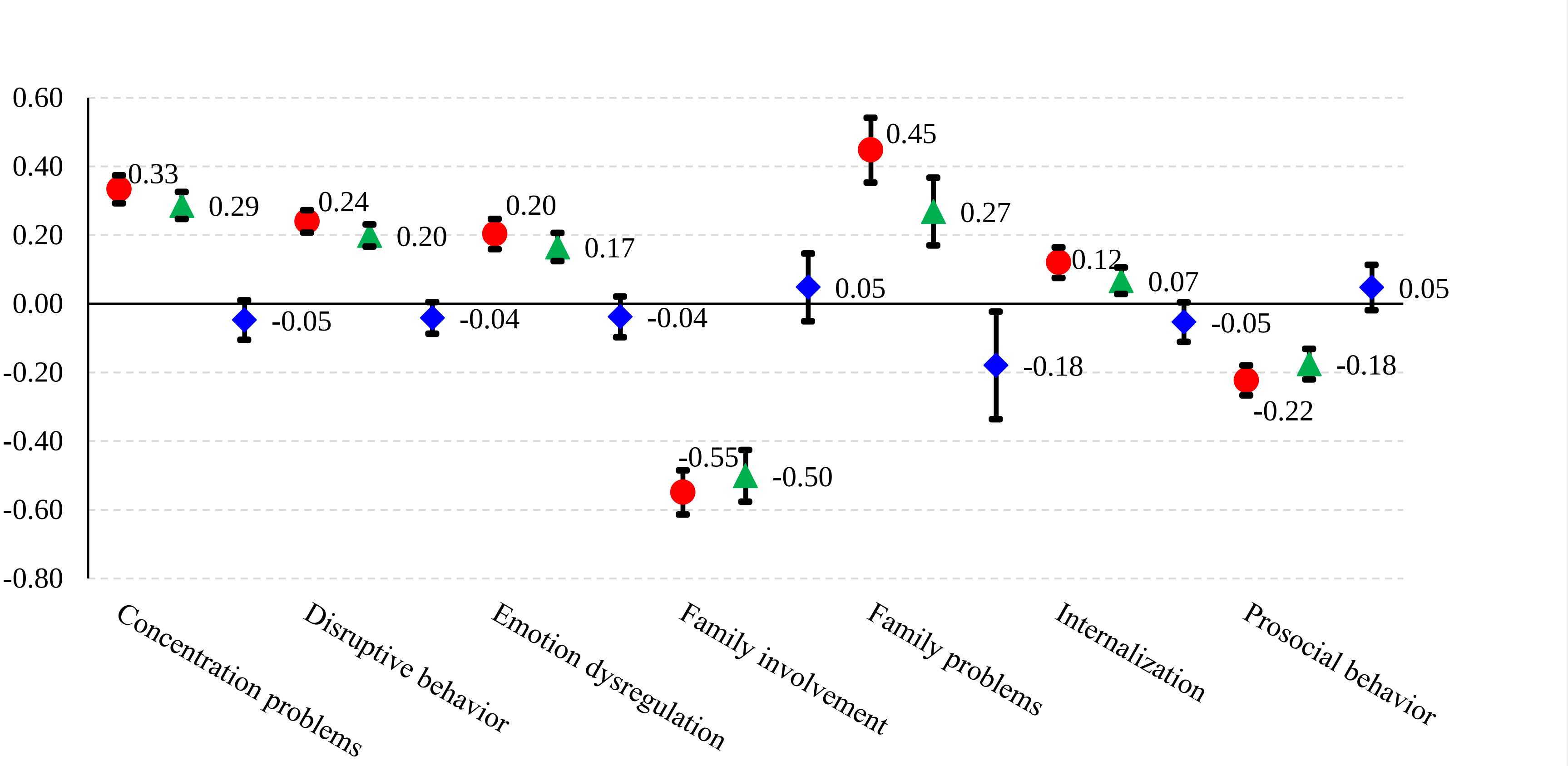
Variance ($\tau^2_{11|T}$) and Heterogeneity (ω_x) of Disparity across Schools

Predictor	Outcome	Estimate ($\hat{\tau}^2_{11|T}$)	SE	*P* value	95% CI		ω_x	95% CI of ω_x		Total Variance ($\tau^2_{00} + \sigma^2$)
Female vs. Male	Concentration problems	0.087	0.009	<0.0001	0.071	0.108	0.242	0.219	0.270	1.480
	Disruptive behavior	0.004	0.001	0.0002	0.002	0.008	0.075	0.059	0.102	0.722
	Emotion dysregulation	0.011	0.004	0.0021	0.006	0.026	0.100	0.074	0.151	1.136
	Family involvement	0.003	0.003	0.2074	0.001	0.565	0.039	0.018	0.544	1.906
	Family problems	0.010	0.003	<0.0001	0.007	0.018	0.149	0.120	0.196	0.473
	Internalization	0.002	0.001	0.0506	0.001	0.013	0.059	0.038	0.139	0.653
	Prosocial behavior	0.010	0.002	<0.0001	0.007	0.015	0.117	0.101	0.141	0.739
Eligible vs. Ineligible for free/reduced price lunch	Concentration problems	0.038	0.007	<0.0001	0.027	0.058	0.160	0.134	0.198	1.480
	Disruptive behavior	0.008	0.002	0.0001	0.005	0.014	0.105	0.084	0.140	0.729
	Emotion dysregulation	0.013	0.005	0.0022	0.007	0.031	0.108	0.081	0.164	1.143
	Family involvement	0.102	0.021	<0.0001	0.070	0.159	0.231	0.192	0.288	1.910
	Family problems	0.028	0.006	<0.0001	0.019	0.045	0.239	0.196	0.306	0.484
	Internalization	0.007	0.003	0.0024	0.004	0.017	0.104	0.078	0.159	0.659
	Prosocial behavior	0.018	0.003	<0.0001	0.013	0.026	0.162	0.138	0.197	0.683
White vs. Black	Concentration problems	0.034	0.007	<0.0001	0.023	0.055	0.152	0.126	0.193	1.480
	Disruptive behavior	0.013	0.003	<0.0001	0.009	0.021	0.135	0.111	0.171	0.734
	Emotion dysregulation	0.029	0.008	0.0002	0.018	0.054	0.159	0.125	0.217	1.152
	Family involvement	0.112	0.027	<0.0001	0.073	0.190	0.241	0.196	0.315	1.914
	Family problems	0.029	0.007	<0.0001	0.019	0.049	0.243	0.197	0.316	0.486
	Internalization	0.003	0.003	0.1284	0.001	0.059	0.068	0.037	0.299	0.657
	Prosocial behavior	0.013	0.003	<0.0001	0.008	0.022	0.130	0.105	0.172	0.755
White vs. Hispanic	Concentration problems	0.036	0.012	0.0013	0.021	0.079	0.160	0.121	0.236	1.421
	Disruptive behavior	0.012	0.004	0.0032	0.007	0.029	0.147	0.108	0.228	0.565
	Emotion dysregulation	0.017	0.009	0.0311	0.007	0.069	0.128	0.084	0.261	1.011
	Family involvement	0.043	0.019	0.0108	0.021	0.128	0.160	0.113	0.276	1.683
	Family problems	0.004	0.004	0.1524	0.001	0.150	0.102	0.054	0.595	0.423

Note : $\omega_x = \sqrt{\tau^2_{11|T}/(\tau^2_{00} + \sigma^2)}$

Figure 1

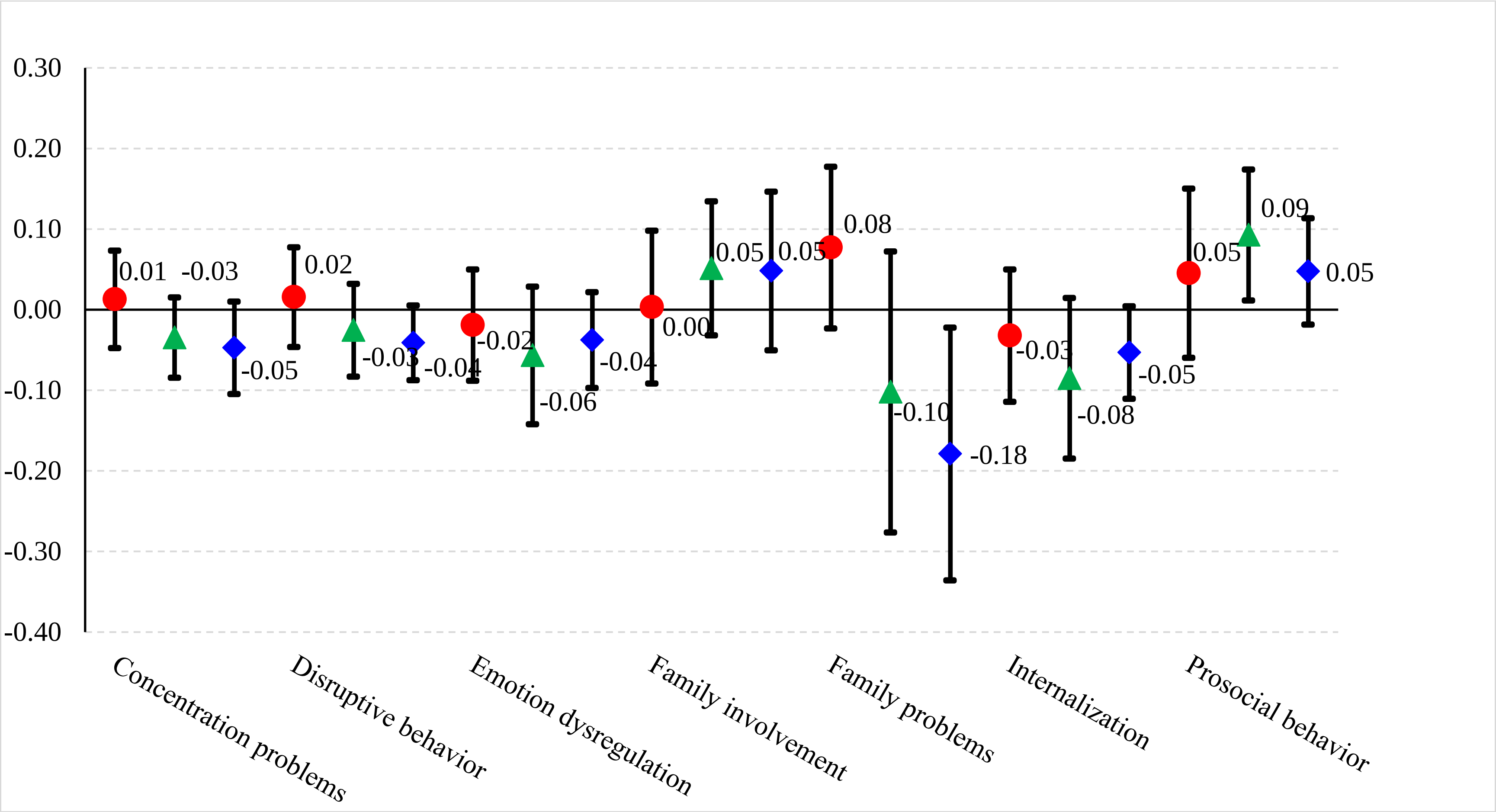
Effect Sizes and 95% CIs of Disparities between Students Eligible and Ineligible for Free/Reduced Price Lunch



Note: Red dots indicate the disparities in the control group (γ_{10}); Green triangles indicate the disparities in the treatment group ($\gamma_{10} + \gamma_{11}$); Blue diamonds indicate the differences in the disparities between the treatment and control groups (γ_{11}).

Figure 2

Effect Sizes and 95% CIs of Interventions on Eligible and Ineligible Free/Reduced Price Lunch



Note: Red dots indicate treatment effects on the ineligible free/reduced price lunch (FRPL); Green triangles indicate treatment effects on the eligible FRPL; Blue diamonds indicate treatment effect differences between eligible and ineligible FRPL (moderated treatment effects).

Supplemental Material - Online Resource

Tables and Figures

Table S1: Two-level HLM Result Summary Regarding Gender (Female vs. Male)

Outcome	Parameters of Interest	Estimate	SE	*P* value	95% CI		Effect Size (ES)	ES 95% CI		Unconditional ICC	Number of Students	Number of Schools
Concentration problems	Disparity in Controls (γ_{10})	-0.52	0.03	<0.0001	-0.57	-0.47	-0.43	-0.47	-0.39	0.03	90,880	387
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.46	0.03	<0.0001	-0.51	-0.41	-0.38	-0.42	-0.34			
	Treatment Effect on Males (γ_{01})	-0.08	0.03	0.0171	-0.15	-0.01	-0.07	-0.12	-0.01			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	-0.02	0.03	0.5151	-0.09	0.04	-0.02	-0.07	0.04			
	Moderated Treatment Effect (γ_{11})	0.06	0.04	0.1045	-0.01	0.13	0.05	-0.01	0.11			
Aggressive/ Disruptive behavior	Disparity in Controls (γ_{10})	-0.30	0.01	<0.0001	-0.32	-0.29	-0.36	-0.38	-0.34	0.06	90,865	387
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.33	0.01	<0.0001	-0.35	-0.31	-0.39	-0.41	-0.36			
	Treatment Effect on Males (γ_{01})	-0.01	0.03	0.8403	-0.05	0.04	-0.01	-0.06	0.05			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	-0.03	0.02	0.1770	-0.07	0.01	-0.03	-0.08	0.02			
	Moderated Treatment Effect (γ_{11})	-0.02	0.01	0.0603	-0.05	0.00	-0.03	-0.06	0.00			
Emotion Dysregulation	Disparity in Controls (γ_{10})	-0.43	0.02	<0.0001	-0.46	-0.39	-0.40	-0.43	-0.37	0.05	37,243	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.42	0.02	<0.0001	-0.46	-0.38	-0.39	-0.43	-0.35			
	Treatment Effect on Males (γ_{01})	-0.07	0.05	0.1459	-0.16	0.02	-0.06	-0.15	0.02			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	-0.06	0.04	0.1396	-0.14	0.02	-0.06	-0.13	0.02			
	Moderated Treatment Effect (γ_{11})	0.01	0.03	0.7871	-0.04	0.06	0.01	-0.04	0.06			
Family involvement	Disparity in Controls (γ_{10})	0.09	0.02	<0.0001	0.05	0.14	0.07	0.04	0.10	0.10	36,998	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	0.10	0.02	<0.0001	0.06	0.14	0.07	0.04	0.10			
	Treatment Effect on Males (γ_{01})	0.07	0.06	0.1948	-0.04	0.18	0.05	-0.03	0.13			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	0.08	0.06	0.1909	-0.04	0.19	0.06	-0.03	0.14			
	Moderated Treatment Effect (γ_{11})	0.01	0.03	0.8384	-0.05	0.06	0.00	-0.04	0.05			
Family problems	Disparity in Controls (γ_{10})	-0.08	0.02	<0.0001	-0.12	-0.05	-0.12	-0.17	-0.07	0.03	35,684	109
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.09	0.02	<0.0001	-0.13	-0.06	-0.14	-0.19	-0.08			
	Treatment Effect on Males (γ_{01})	-0.01	0.05	0.8609	-0.11	0.09	-0.01	-0.16	0.14			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	-0.02	0.05	0.6897	-0.12	0.08	-0.03	-0.17	0.11			
	Moderated Treatment Effect (γ_{11})	-0.01	0.03	0.6725	-0.06	0.04	-0.02	-0.09	0.06			
Internalization	Disparity in Controls (γ_{10})	-0.09	0.01	<0.0001	-0.11	-0.06	-0.11	-0.14	-0.07	0.05	37,243	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.09	0.01	<0.0001	-0.11	-0.06	-0.11	-0.14	-0.08			
	Treatment Effect on Males (γ_{01})	-0.07	0.04	0.0496	-0.15	0.00	-0.09	-0.18	0.00			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	-0.08	0.04	0.0410	-0.15	0.00	-0.09	-0.18	0.00			
	Moderated Treatment Effect (γ_{11})	0.00	0.02	0.9220	-0.04	0.03	0.00	-0.04	0.04			
Prosocial behaviors	Disparity in Controls (γ_{10})	0.25	0.01	<0.0001	0.23	0.28	0.30	0.27	0.33	0.06	83,679	328
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	0.27	0.01	<0.0001	0.25	0.29	0.31	0.29	0.34			
	Treatment Effect on Males (γ_{01})	0.06	0.04	0.1459	-0.02	0.13	0.06	-0.02	0.15			
	Treatment Effect on Females ($\gamma_{01} + \gamma_{11}$)	0.07	0.03	0.0169	0.01	0.13	0.08	0.01	0.15			
	Moderated Treatment Effect (γ_{11})	0.01	0.02	0.4295	-0.02	0.05	0.02	-0.02	0.06			

Table S2: Two-level HLM Result Summary Regarding Race (White vs. Black)

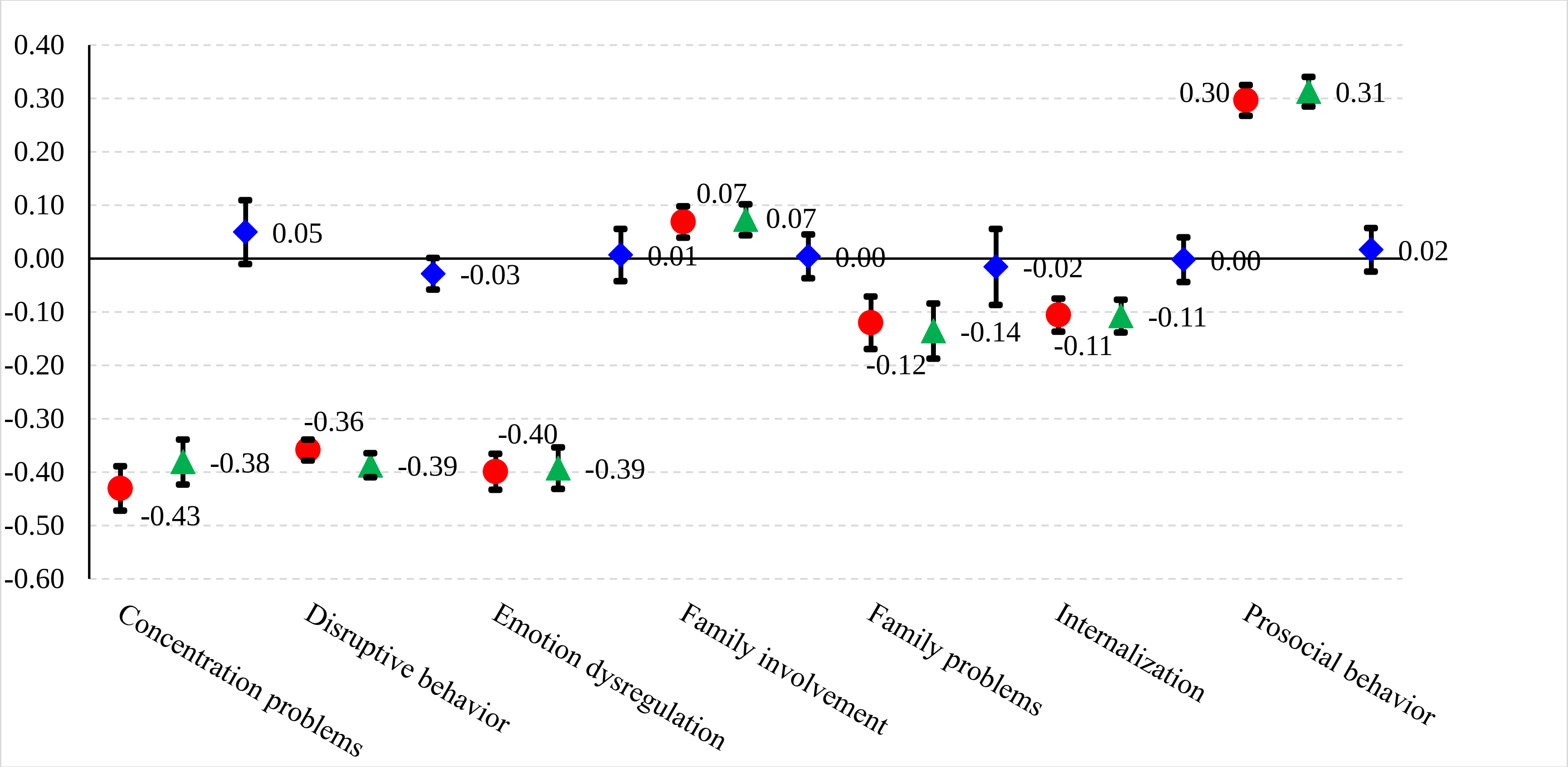
Outcome	Parameters of Interest	Estimate	SE	*P* value	95% CI		Effect Size (ES)	ES 95% CI		Unconditional ICC	Number of Students	Number of Schools
Concentration problems	Disparity in Controls (γ_{10})	-0.32	0.02	<0.0001	-0.36	-0.27	-0.26	-0.30	-0.22	0.03	80,401	387
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.31	0.02	<0.0001	-0.36	-0.27	-0.26	-0.30	-0.22			
	Treatment Effect on Black (γ_{01})	-0.05	0.03	0.0944	-0.12	0.01	-0.04	-0.10	0.01			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	-0.05	0.03	0.1036	-0.11	0.01	-0.04	-0.09	0.01			
	Moderated Treatment Effect (γ_{11})	0.00	0.03	0.9184	-0.06	0.07	0.00	-0.05	0.06			
Aggressive/ Disruptive behavior	Disparity in Controls (γ_{10})	-0.31	0.02	<0.0001	-0.34	-0.28	-0.36	-0.39	-0.32	0.06	80,387	387
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.28	0.02	<0.0001	-0.32	-0.25	-0.33	-0.37	-0.30			
	Treatment Effect on Black (γ_{01})	-0.03	0.03	0.1736	-0.08	0.02	-0.04	-0.10	0.02			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	-0.01	0.02	0.5586	-0.05	0.03	-0.01	-0.06	0.03			
	Moderated Treatment Effect (γ_{11})	0.02	0.02	0.3217	-0.02	0.07	0.03	-0.03	0.08			
Emotion Dysregulation	Disparity in Controls (γ_{10})	-0.28	0.03	<0.0001	-0.34	-0.22	-0.26	-0.31	-0.20	0.05	32,061	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.23	0.03	<0.0001	-0.30	-0.16	-0.22	-0.28	-0.15			
	Treatment Effect on Black (γ_{01})	-0.09	0.05	0.0585	-0.19	0.00	-0.09	-0.17	0.00			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	-0.05	0.04	0.2506	-0.13	0.03	-0.04	-0.12	0.03			
	Moderated Treatment Effect (γ_{11})	0.05	0.05	0.3276	-0.05	0.14	0.04	-0.04	0.13			
Family involvement	Disparity in Controls (γ_{10})	0.53	0.04	<0.0001	0.45	0.61	0.38	0.33	0.44	0.11	31,837	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	0.62	0.06	<0.0001	0.50	0.74	0.45	0.36	0.54			
	Treatment Effect on Black (γ_{01})	0.04	0.06	0.5049	-0.08	0.15	0.03	-0.05	0.11			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	0.13	0.07	0.0511	0.00	0.26	0.09	0.00	0.18			
	Moderated Treatment Effect (γ_{11})	0.09	0.07	0.1899	-0.04	0.22	0.06	-0.03	0.16			
Family problems	Disparity in Controls (γ_{10})	-0.09	0.03	0.0038	-0.15	-0.03	-0.13	-0.21	-0.04	0.03	30,926	109
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	-0.13	0.04	0.0028	-0.22	-0.05	-0.19	-0.31	-0.06			
	Treatment Effect on Black (γ_{01})	0.00	0.06	0.9936	-0.11	0.11	0.00	-0.16	0.16			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	-0.04	0.06	0.4781	-0.16	0.07	-0.06	-0.22	0.11			
	Moderated Treatment Effect (γ_{11})	-0.04	0.06	0.4973	-0.16	0.08	-0.06	-0.23	0.11			
Internalization	Disparity in Controls (γ_{10})	-0.02	0.02	0.3545	-0.06	0.02	-0.02	-0.07	0.03	0.05	32,061	180
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	0.02	0.02	0.2077	-0.01	0.06	0.03	-0.02	0.08			
	Treatment Effect on Black (γ_{01})	-0.08	0.04	0.0480	-0.16	0.00	-0.10	-0.20	0.00			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	-0.04	0.04	0.2682	-0.11	0.03	-0.05	-0.14	0.04			
	Moderated Treatment Effect (γ_{11})	0.04	0.03	0.1399	-0.01	0.10	0.05	-0.02	0.12			
Prosocial behaviors	Disparity in Controls (γ_{10})	0.22	0.02	<0.0001	0.18	0.26	0.26	0.21	0.30	0.06	74,135	328
	Disparity in Treated ($\gamma_{10} + \gamma_{11}$)	0.19	0.02	<0.0001	0.15	0.22	0.21	0.18	0.25			
	Treatment Effect on Black (γ_{01})	0.09	0.03	0.0061	0.02	0.15	0.10	0.03	0.17			
	Treatment Effect on White ($\gamma_{01} + \gamma_{11}$)	0.05	0.03	0.1173	-0.01	0.12	0.06	-0.01	0.13			
	Moderated Treatment Effect (γ_{11})	-0.04	0.03	0.1790	-0.09	0.02	-0.04	-0.10	0.02			

Table S3: Two-level HLM Result Summary Regarding Race (White vs. Hispanic)

Outcome	Parameters of Interest	Estimate	SE	*P* value	95% CI		Effect Size (ES)	ES 95% CI		Unconditional ICC	Number of Students	Number of Schools
Concentration problems	Disparity in Controls (γ_{10})	0.01	0.03	0.6241	-0.04	0.07	0.01	-0.04	0.06	0.03	41,820	346
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.02	0.04	0.6086	-0.05	0.09	0.02	-0.04	0.08			
	Treatment Effect on Hispanic (γ_{01})	-0.06	0.04	0.1431	-0.15	0.02	-0.05	-0.13	0.02			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	-0.06	0.03	0.0567	-0.12	0.00	-0.05	-0.10	0.00			
	Moderated Treatment Effect (γ_{11})	0.00	0.05	0.9266	-0.09	0.10	0.00	-0.07	0.08			
Aggressive/ Disruptive behavior	Disparity in Controls (γ_{10})	0.07	0.02	0.0010	0.03	0.11	0.09	0.04	0.15	0.04	41,819	346
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.10	0.02	<0.0001	0.05	0.14	0.13	0.07	0.19			
	Treatment Effect on Hispanic (γ_{01})	-0.04	0.03	0.2050	-0.11	0.02	-0.06	-0.14	0.03			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	-0.02	0.02	0.5060	-0.06	0.03	-0.02	-0.08	0.04			
	Moderated Treatment Effect (γ_{11})	0.03	0.03	0.4137	-0.04	0.09	0.03	-0.05	0.12			
Emotion Dysregulation	Disparity in Controls (γ_{10})	0.22	0.04	<0.0001	0.14	0.29	0.21	0.14	0.29	0.03	15,168	147
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.19	0.04	<0.0001	0.11	0.27	0.19	0.11	0.27			
	Treatment Effect on Hispanic (γ_{01})	-0.01	0.06	0.9065	-0.13	0.12	-0.01	-0.13	0.12			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	-0.03	0.05	0.5054	-0.13	0.06	-0.03	-0.13	0.06			
	Moderated Treatment Effect (γ_{11})	-0.03	0.06	0.6632	-0.14	0.09	-0.03	-0.14	0.09			
Family involvement	Disparity in Controls (γ_{10})	0.47	0.04	<0.0001	0.39	0.56	0.36	0.30	0.43	0.08	15,041	147
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.52	0.07	<0.0001	0.39	0.65	0.40	0.30	0.50			
	Treatment Effect on Hispanic (γ_{01})	0.07	0.08	0.4030	-0.09	0.23	0.05	-0.07	0.18			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	0.12	0.07	0.0940	-0.02	0.25	0.09	-0.02	0.19			
	Moderated Treatment Effect (γ_{11})	0.05	0.08	0.5479	-0.11	0.20	0.04	-0.08	0.15			
Family problems	Disparity in Controls (γ_{10})	0.06	0.03	0.0352	0.00	0.11	0.09	0.01	0.16	0.03	14,729	104
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.04	0.03	0.2782	-0.03	0.10	0.06	-0.05	0.16			
	Treatment Effect on Hispanic (γ_{01})	-0.02	0.04	0.5862	-0.09	0.05	-0.03	-0.14	0.08			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	-0.04	0.06	0.5223	-0.16	0.08	-0.06	-0.24	0.12			
	Moderated Treatment Effect (γ_{11})	-0.02	0.05	0.7086	-0.12	0.08	-0.03	-0.18	0.12			
Internalization	Disparity in Controls (γ_{10})	0.12	0.02	<0.0001	0.07	0.17	0.15	0.09	0.21	0.03	15,168	147
	Disparity in Treated ($\gamma_{10}+\gamma_{11}$)	0.15	0.03	<0.0001	0.09	0.20	0.18	0.12	0.24			
	Treatment Effect on Hispanic (γ_{01})	-0.06	0.04	0.1517	-0.14	0.02	-0.07	-0.17	0.03			
	Treatment Effect on White ($\gamma_{01}+\gamma_{11}$)	-0.03	0.04	0.3749	-0.10	0.04	-0.04	-0.13	0.05			
	Treatment Effect on Hispanic (γ_{01})	0.03	0.04	0.4604	-0.04	0.10	0.03	-0.05	0.12			

Figure S1

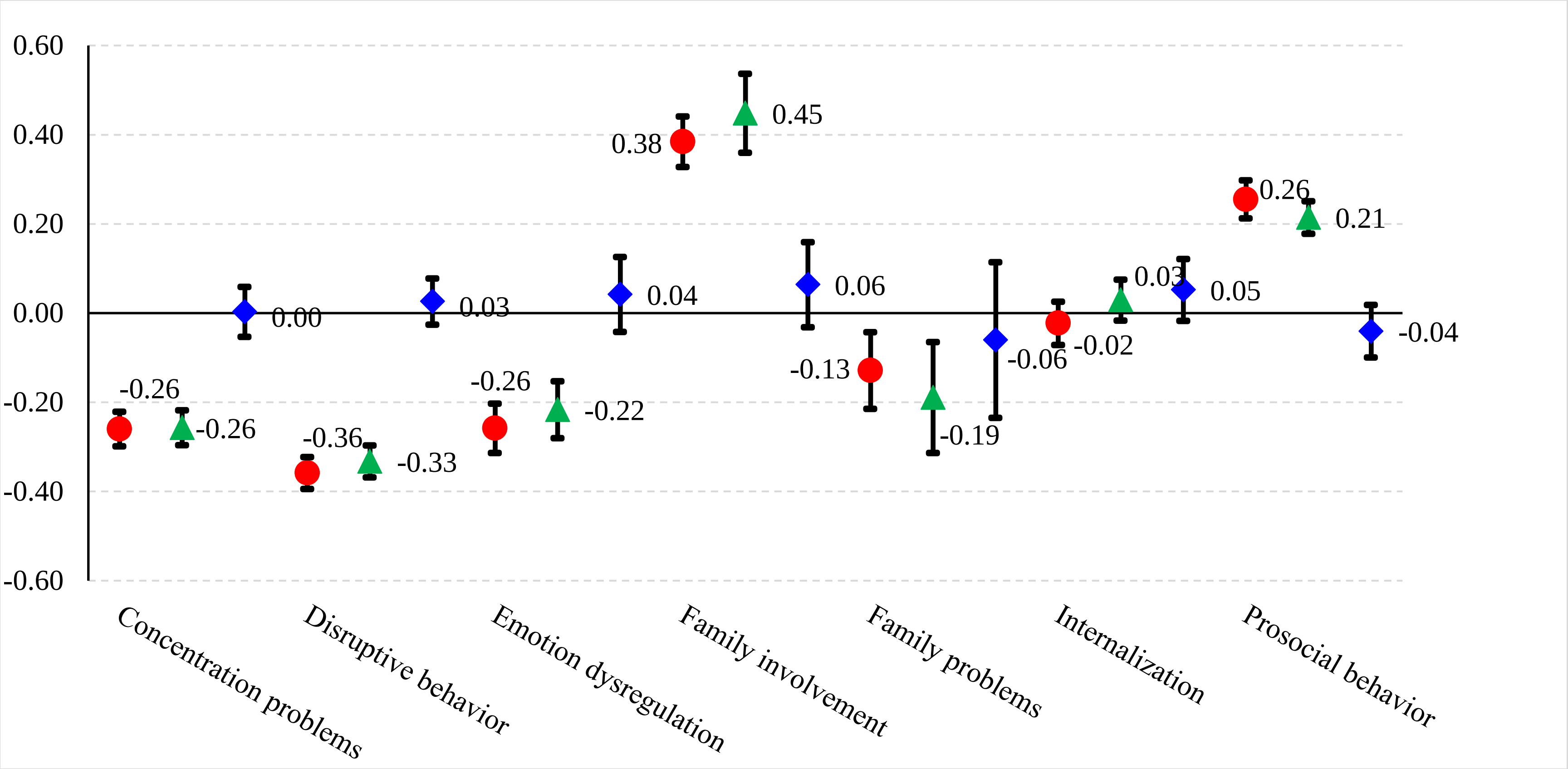
Effect Sizes and 95% CIs of Disparities between Females and Males



Note: Red dots indicate the disparities in the control group (γ_{10}); Green triangles indicate the disparities in the treatment group ($\gamma_{10} + \gamma_{11}$); Blue diamonds indicate the differences in the disparities between the treatment and control groups (γ_{11}).

Figure S2

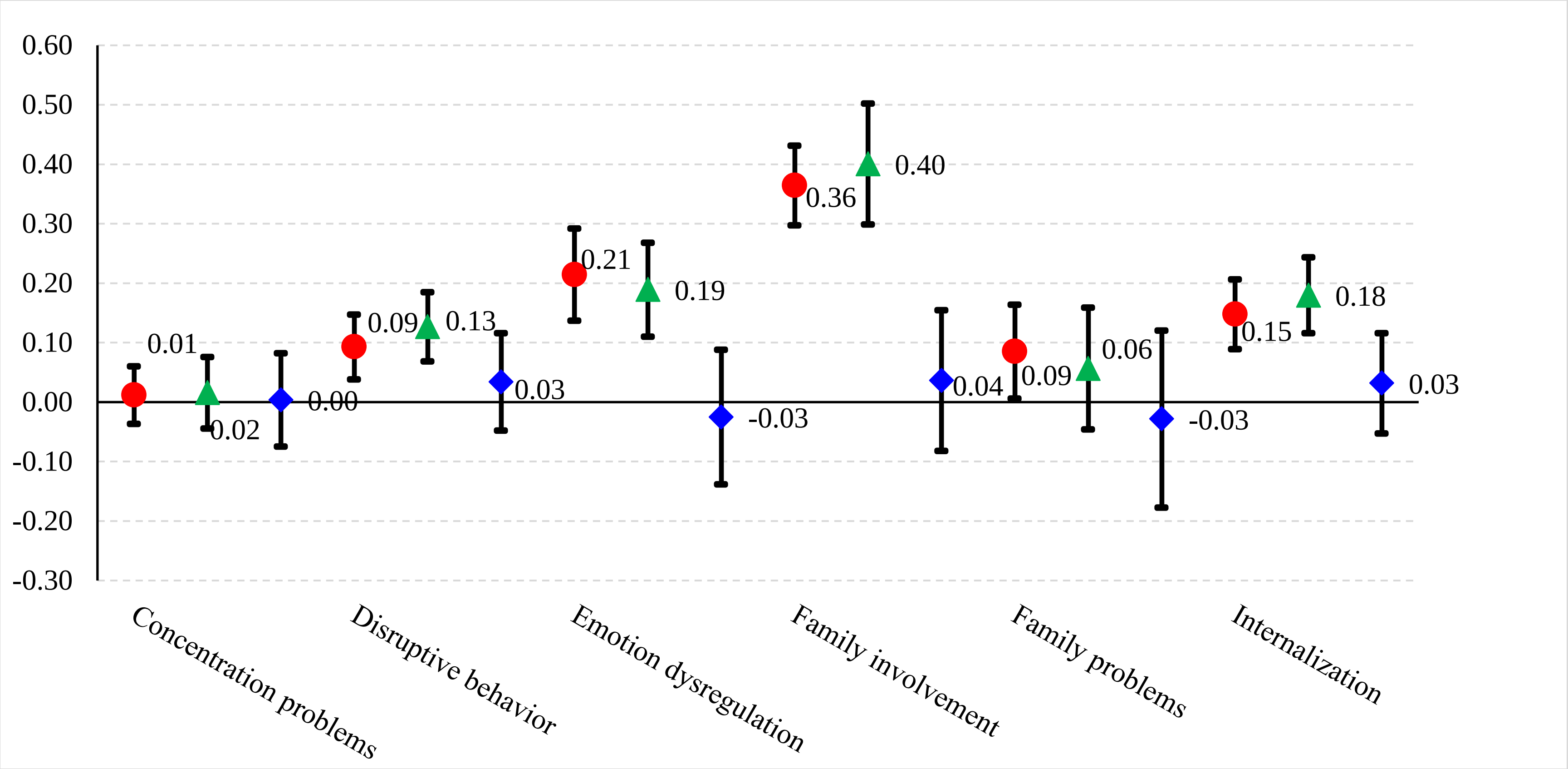
Effect Sizes and 95% CIs of Disparities between White and Black



Note: Red dots indicate the disparities in the control group (γ_{10}); Green triangles indicate the disparities in the treatment group ($\gamma_{10} + \gamma_{11}$); Blue diamonds indicate the differences in the disparities between the treatment and control groups (γ_{11}).

Figure S3

Effect Sizes and 95% CIs of Disparities between White and Hispanic



Note: Red dots indicate the disparities in the control group (γ_{10}); Green triangles indicate the disparities in the treatment group ($\gamma_{10} + \gamma_{11}$); Blue diamonds indicate the differences in the disparities between the treatment and control groups (γ_{11}).

Figure S4

Heterogeneity of Disparity across Schools and 95% CI between Students Eligible and Ineligible for Free/Reduced Price Lunch

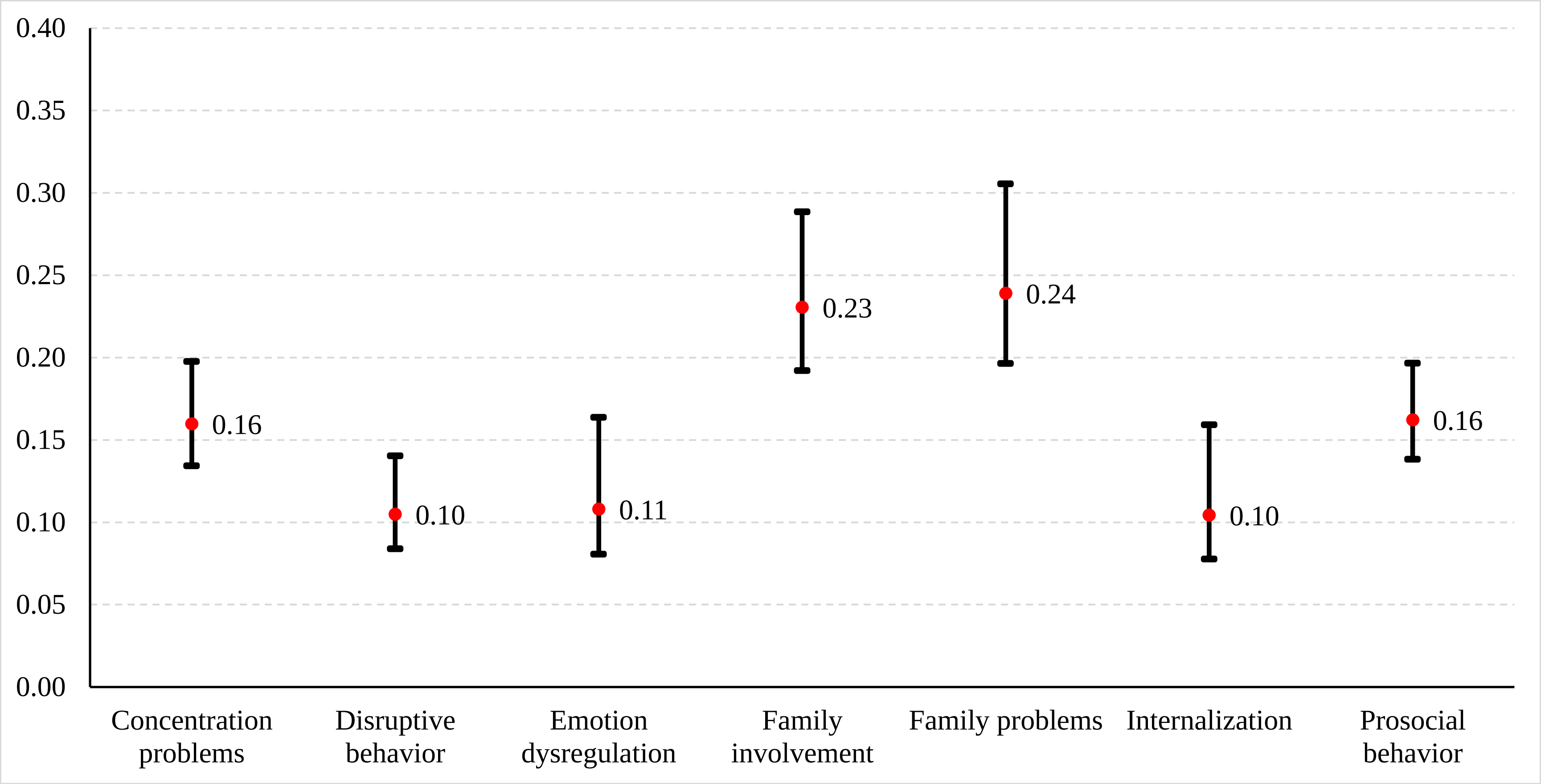
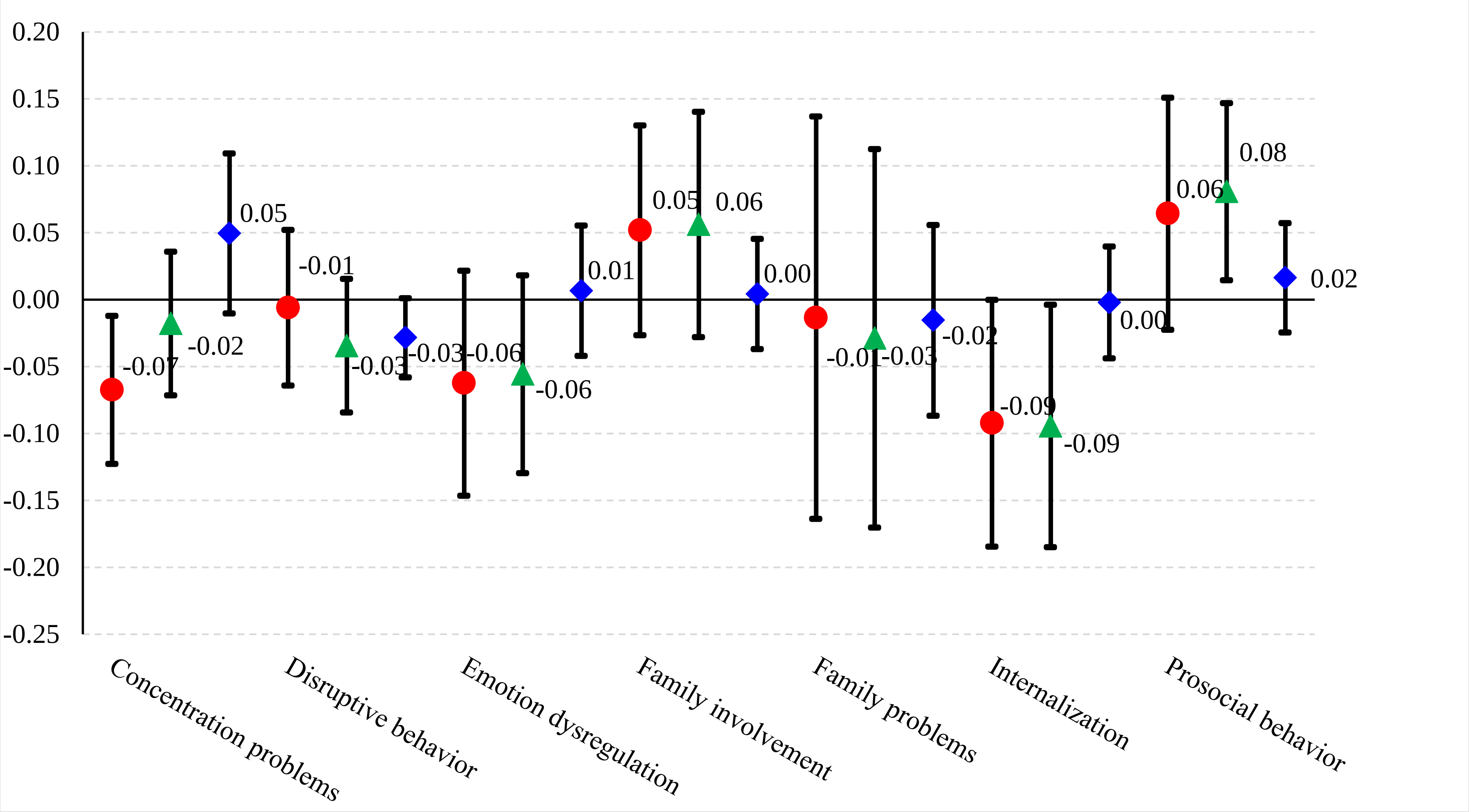


Figure S5

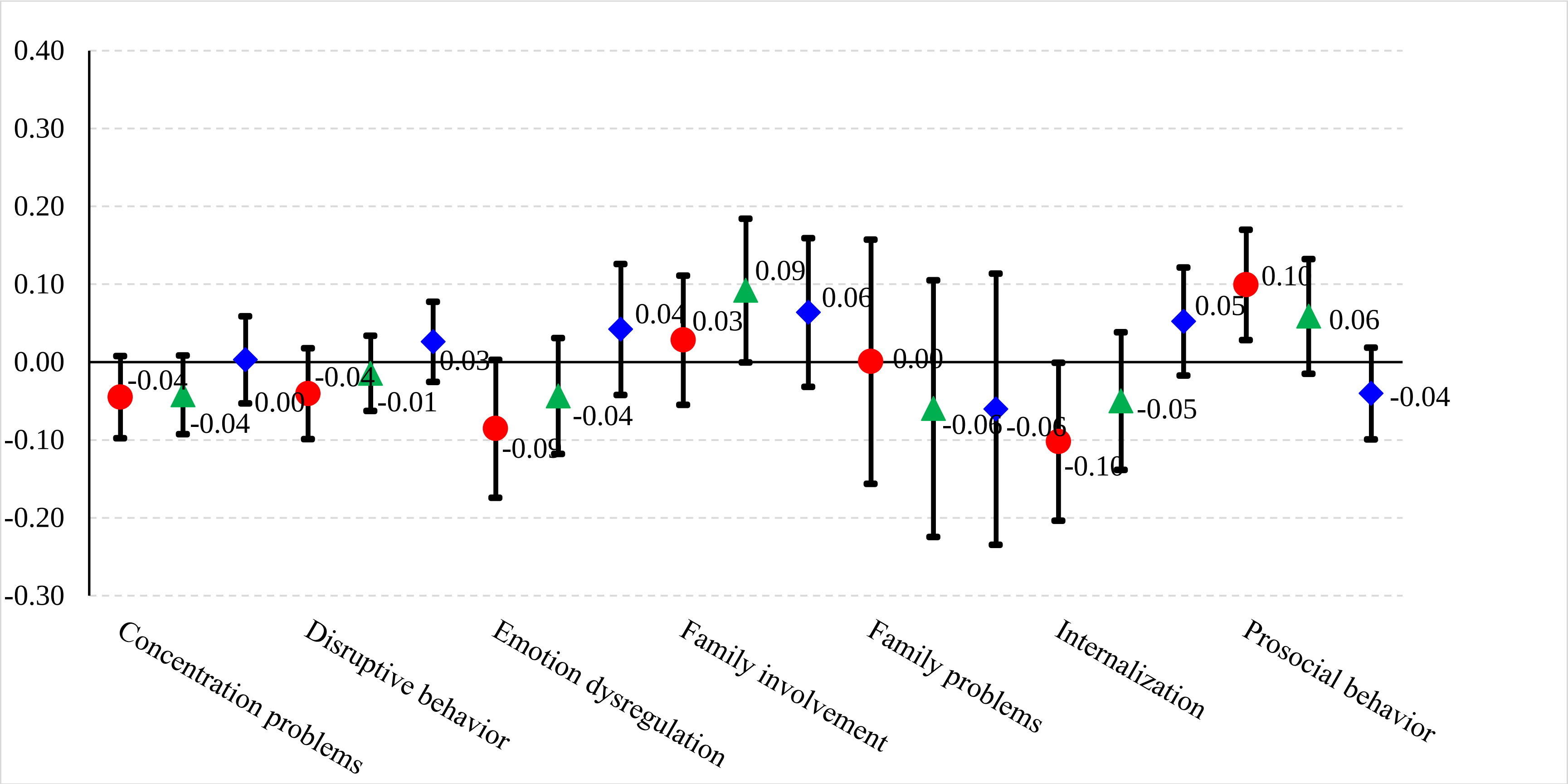
Effect Sizes and 95% CIs of Interventions on Females and Males



Note: Red dots indicate treatment effects on Males; Green triangles indicate treatment effects on Females; Blue diamonds indicate treatment effect differences between Females and Males (moderated treatment effects).

Figure S6

Effect Sizes and 95% CIs of Interventions on White and Black



Note: Red dots indicate treatment effects on Black; Green triangles indicate treatment effects on White; Blue diamonds indicate treatment effect differences between White and Black (moderated treatment effects).