Running Head: HISTORICAL EVENTS AND RCTS

Accounting for Traumatic Historical Events in

Randomized Controlled Trials

Abstract

As an example of how historical events may influence the findings and interpretations of a randomized trial, we use a school-based evaluation of a classroom management program that our group conducted in a nearby district before and after the shooting of Michael Brown in Ferguson, Missouri. The findings suggest that the event differentially affected teacher and student response within and across conditions. Black teachers benefitted more from the intervention compared to White teachers as evidenced by their independently observed classroom management skills and praise-to-reprimand ratios; however, these effects were minimized or disappeared after the event. Additionally, although the intervention equally benefitted the academic achievement of Black and White students before the event, the Black-White achievement gap widened after the event. Implications for the design, analysis, and reporting of findings from randomized controlled trials are discussed.

Keywords: history, achievement gap, classroom management, academic achievement

Public Health Significance

This cluster randomized trial conducted before and after the Michael Brown shooting revealed that historical events can differentially effect participants both within and across conditions. Although the intervention equally benefitted the academic achievement of Black and White students before the event, the Black-White achievement gap widened after the event. Findings suggest that the Black-White achievement gap will not be reduced by schools alone and that human subject scholars need to examine the influence of traumatic historical events on study findings and interpretations.

Accounting for Traumatic Historical Events in Educational Randomized Controlled Trials

The international crises related to the COVID-19 pandemic and police brutality have highlighted the impact of historical events on the conduct of scientific studies with human subjects. In the US, many, if not all psychological and educational studies were interrupted as schools closed throughout the country in the spring of 2020 and social distancing measures were enacted. This created obvious challenges for research designs, particularly randomized trials, that were left without post-intervention scores on primary outcome variables. Perhaps less obvious or considered is the impact of these events on the internal and external validity of these studies.

Campbell (1957) identified history, events experienced by participants that influence the dependent variable but are not part of the study design, as one of seven threats to a study's internal validity The randomized experimental design controls for history to the extent that participants in each condition are equally likely to be exposed to and impacted by the historical event. Thus, it is unlikely that any observed differences between groups at post-test are due to the event.

However, even if the historical event equally affects participants in both comparison groups, it may still interfere with study findings and interpretation because the event may affect the quality or dosage of the intervention. For instance, if the event makes it difficult for participants to receive or benefit from the intervention, the event may lower the effect size and lead to a false conclusion that it is unhelpful. While the conclusion that the intervention did not work in the context of the historical event may be accurate, it is possible the intervention might work in the absence of that event. To the extent the event is uncommon, this conclusion would be problematic.

One way to conceptualize how historical events may affect studies is to consider how traumatic the event might be for participants. Several elements have been identified that influence how much an individual may be impacted by a given traumatic event (Vogt et al., 2007). First, the severity and duration of a traumatic event, or its potential to cause death, injury, or disrupted life circumstances, have major roles in how it affects humans who experience it (McLean et al., 2013). Second, an individual's specific demographic background and/or preexisting conditions can also influence responses to traumatic events (Brewin et al., 2000). For instance, an individual who identifies as a Black male may have had a particularly negative response to the murders of George Floyd and Armaud Arbery compared to those of other racial and gender identities. Additionally, someone with an anxiety disorder, may have an escalation of their anxiety and its impact on their functioning in relation to the pandemic relative to others without pre-existing anxiety conditions.

Case Example: An RCT in the midst of the Ferguson, Missouri Protests

As an example of how historical events may influence the findings and interpretations of a randomized controlled trial (RCT), we use a school-based educational trial that our group conducted before and after the shooting of Michael Brown in Ferguson, Missouri. On August 9, 2014 a police officer shot and killed Michael Brown, an unarmed Black high school student, an event that sparked widescale protests and rioting in the surrounding communities (Kochel, 2015). Nearby school districts delayed the opening of school because of safety concerns for their students and faculty. Protests continued through the Fall semester and escalated again after the officer was not indicted for the shooting in November, 2014. Nearby school districts again closed for one or more days in anticipation of the grand jury announcement. Aside from the logistical life disruptions that these events presented to students and teachers in these communities, they were perceived differently by individuals based on their race and ethnicity (Kochel, 2015). In general, Black citizens experience and interpret police brutality more personally and negatively than White citizens (Strickler & Lawson, 2020).

Our RCT took place in one of the nearby school districts; all of the schools in our study were within a 10 mile radius of where the shooting occurred. The purpose of the original study was to evaluate the effects of a classroom management program, CHAMPS, on middle school teacher and student outcomes (Sprick et al., 2009). The study design involved recruiting four annual cohorts of teachers, each randomly assigned to treatment or control conditions within each of nine total school buildings. The historical event happened during the fall of the third cohort.

We have previously reported the main effects of this study, which revealed that the CHAMPS training program significantly improved teacher practices and student disruptive behaviors, time-on-task, work completion, problem solving, and communication skills (AUTHOR, 2020). Given recent events, we were curious if we could detect any differential treatment effects on the third cohort of teachers and students that experienced the aftermath of shooting and protests. Although we expected teachers and students in both treatment conditions in cohort three to have comparable levels of severity, proximity, and duration of exposure to the event, we considered whether certain student and teacher demographics and preexisting risk characteristics may have further moderated the impact of the event on study outcomes. For instance, the majority of students in the schools were Black and thus may have been particularly affected by the event. Additionally, we collected baseline measures of teacher stress and student depressive symptoms that may have served as risk factors for more severe reactions to the traumatic event.

The current study examined cohort-specific main and moderated effects of the CHAMPS intervention during a major and potentially traumatic event that occurred during the third year of the trial. We began with six research questions:

Research Question (RQ) 1: Did the Michael Brown shooting and subsequent protests during the third cohort of the study alter the outcomes on teachers or students? (Main effect of the event.)

RQ 2: Did the Michael Brown shooting and subsequent protests during the third cohort of the study alter CHAMPS intervention effects on teachers or students? (Event as a moderator of intervention.)

RQ 3: Did Black students or teachers have disproportionate response to the shooting and protests compared to White students and teachers? (Race as a moderator of the event.)

RQ 4: Did Black students or teachers have different treatment effects compared to White students and teachers? (Race as a moderator of the intervention.)

RQ 5: Did Black students or teachers receive different treatment effects compared to White students and teachers depending on the event? (Race and event as moderators of the intervention.)

RQ 6: Did teachers with higher levels of baseline distress have a disproportionate adverse treatment response after the event compared to those with more normative levels of stress or depressive symptoms? (Baseline measures as moderators of the event.)

Method

Participants

Middle school teacher and student participants were recruited from an urban school district in the Midwest U.S. Participants were recruited as part of a group RCT of the CHAMPS behavior management and coaching program. Eligible teacher participants included sixth- to eighth- grade English or Math teachers who consented to participate in the project. Parent consent and student assent were obtained for student participants recruited from classrooms of participating teachers.

A final teacher sample of 102 and a student sample of 1,450 agreed to participate in the study across all four cohorts. Our analytic sample included 86 teachers and 1,069 students in Cohorts 1-3; we excluded cohort 4 from the current analyses because students and teachers in this cohort were from a different district. A cluster random assignment design was utilized. Teachers were randomly assigned to receive CHAMPS or to a wait-list, business as usual control group within school, with the constraint that the number of intervention teachers to be no more than one more or less than the number of control teachers. Teacher participants were recruited and randomized across three cohorts [year 1: 26 teachers (13 intervention), 394 students; year 2: 36 teachers (18 intervention), 382 students; year 3: 24 teachers (12 intervention), 293 students]. Cohorts 1 & 2 were before the Michael Brown event and Cohort 3 was immediately after the Michael Brown event.

Student participants were 49.6% female and 74.5% African American, 21.1% White, and 4.4% other race (Hispanic/Latinx, Asian, etc.). The percentage of students in sixth, seventh, and eighth grade was equal to 42.2%, 33.6%, and 24.2%, respectively. Overall, 62.5% of students qualified for free/reduced-priced lunch, and 8.4% of the sample received special education services. Teacher participants were 79.1% female and 70.9% White, 25.6% African American, and 4.5% other. Teachers' ages ranged from 23 to 63 years (M = 37.8, SD = 8.8), whereas teaching experience ranged from 1.0 to 23.0 (M = 10.4, SD = 6.3).

Procedures

The study had high rates of enrollment for eligible teachers (73%) and students (75%). University Institutional Review Board and the participating school district approved the study protocol. Teachers and students were recruited at the beginning of the school year. Data were collected at the beginning of the school year, prior to the intervention, and at the end of the school year, post-intervention. All pre-intervention assessments occurred in mid-September to mid-October. Post-intervention assessments were collected in late April and May of the same academic year. Observations were also collected at baseline (Time 1) and three times following intervention: November (Time 2), February (Time 3), and April/May (Time 4).

CHAMPS Training. Intervention teachers received training and coaching to deliver CHAMPS (Sprick et al., 2009). In three sequential, annual cohorts of between 8 to 18 teachers in the CHAMPS condition attended three full-day group trainings, back-to-back sessions in late-October and an additional session in late-November/early-December. All trainings were facilitated by a certified CHAMPS trainer supervised by the program developer. Additionally, an on-site doctoral-level coach who was trained and supervised by the program developer supported teacher implementation following sessions.

CHAMPS is a comprehensive curriculum for improving teacher classroom management and relationship skills. The CHAMPS model targets teachers' use of effective classroom management strategies by promoting positive relationships with all students and by strengthening the relevance and engagement of instruction. The key principles for an organized and effective classroom are summarized by the acronym STOIC mentioned previously: Structure classroom, Teach expectations, Observe and supervise, Interact positively, and Correct fluently. The training and subsequent coaching support focuses on building teacher competence in each of these five domains. Training occurs in seven modules: developing a vision, organization, developing and teaching expectations, proactive teaching, student motivation, data-based decisions, and calm and consistent corrections. CHAMPS includes a host of well-developed and user-friendly materials to support teacher implementation of the practices. These include the companion books, *CHAMPS: A Proactive and Positive Approach to Classroom Management* and the *Teacher's Encyclopedia of Behavior Management: 100 Problems/500 Plans*; CHAMPS Teacher Planners for keeping on track with the approach; and the Making Every Second Count DVD series.

CHAMPS Coaching. In this study, the CHAMPS coach was a doctoral-level special educator. The coaching model is manualized, partnership-oriented, and involves giving teachers ongoing explicit feedback about their implementation. In between each workshop session, the CHAMPS coach observed the teachers in the classroom and met with them individually for up to one hour every week. We defined a minimal dose that each teacher needed to receive as a total of four visits with the coach. The first visit focused on establishing rapport and setting goals. The second visit focused on providing the teacher with explicit feedback based on the coach's classroom observations and developing a plan based on the teacher's own goals. Subsequent visits were tailored to each teacher based on this goal setting and planning. The coach recorded any contact with teachers, including brief check-ins, to reviewing strategies and schedule the next meeting. During the individual coaching sessions, the coach reviewed workshop content and supported goal setting for use of strategies, provided feedback on teacher skills and interpersonal teaching processes, modeled effective practice, role-played potential barriers and challenges, and supported action planning. CHAMPS is a universal intervention for teachers, meaning that the intervention is intended for all teachers regardless of skill level. However, the CHAMPS coach differentiated the amount of coaching provided to teachers based on their need for supports. The mean time spent with a teacher by the coach, outside of classroom observations was 147 minutes (range = 48 to 358 minutes).

Control Condition. Teachers assigned to the wait-list control condition continued their business as usual teaching and professional development opportunities during the study period. Due to the wait-list design, control condition teachers were offered the CHAMPS intervention immediately after their period of participation in the evaluation component of the project ended. Teachers in both conditions were compensated for their time and effort in completing surveys as part of the project.

Measures of Implementation Fidelity and Teacher Practices

Direct Observations. Classroom observations were conducted by independent observers blind to the intervention condition. Classroom-level observations, including measures of teacher implementation fidelity and adherence were collected across four time points. The first observation occurred in October prior to receiving CHAMPS training or coaching. The second observation in November after teachers received workshop sessions 1 and 2 and at least one coaching visit. The third observation occurred in February after all three workshops were completed and the minimal dose of coaching delivered. The final observation occurred at the end of the school year

(April/May). All observations occurred in classrooms during instructional times. The pre and post classroom observations were an aggregation of a series of four 5-minute observations made by the same observer on a single classroom visit, whereas the second and third observations were both 20 minutes in length. Student-level observations were collected on two occasions, at baseline and at the end-of-the-school-year.

Teacher Implementation Fidelity to CHAMPS. Independent observers conducted direct observations of teacher implementation fidelity using the *STOIC Rating Form* across the four timepoints described previously (Sprick, 2013). STOIC provides global ratings of each of the five key domains of CHAMPS practices: Structure classroom, Teach expectations, Observe and supervise, Interact positively, and Correct fluently. Independent observers rate each of these five domains on a 0 (*no evidence*) to 4 (*full evidence*) rating scale, and we computed a summary score of these ratings as a measure of adherence. The STOIC was not gathered at baseline for cohort 1 of the study because the measure was not available at the start of the project, but all other time points were gathered. Analyses examining changes on the STOIC used other similar measures described below to adjust for baseline differences. Prior to data collection, observers attended a two hour training focused on using the STOIC and practiced coding videos of actual classrooms. They were allowed to collect data only after reaching agreement with a master coder. The ICC (One-Way Random Effects Absolute Agreement) for STOIC summary scores ranged from .92 to .97 at each measurement time point.

In addition, we conducted 20-minute classroom observations using the *Classroom Assessment Scoring System-Secondary* (CLASS-S; Pianta et al., 2008) at baseline and across the same direct observation time points as the STOIC. The CLASS-S asks observers to provide global ratings of specific aspects of a classroom's emotional support, organization, and instructional support on a 7-point scale with higher scores indicating more adaptive environments. All observers attended two full day trainings led by a CLASS-S master trainer. They then completed an online coding test of actual classroom interactions and needed to reach a high level of agreement with the CLASS-S master coder before being certified to collect data. Additionally, observers needed to repeat the certification each year of the project. Because we only collected post-intervention STOIC ratings for the first cohort, we used baseline Climate subscale as a covariate to equate classrooms on baseline climate. The CLASS-S scales have been shown to be highly reliable and to predict student achievement and social outcomes in a number of studies of large numbers of 5th graders (NICHHD, 2005) and work with teachers in secondary settings (Allen et al., 2013). The interclass correlation for the Climate subscale across all time periods was .75.

Teacher Use of Proactive Strategies. Independent observers also conducted direct observations of teacher use of proactive strategies with using the *Multi-Option Observation System for Experimental Studies* (MOOSES; Tapp, 2004) interface for hand held computers to gather real time data using the *Brief Classroom Interaction Observation Revised* observation code (BCIO-R; Reinke et al., 2015). These observations occurred at the same timepoints as the STOIC and CLASS-S, but not by the same observer who collected those observations.

The BCIO-R is a 20-minute class-wide observation of the frequency of teacher use of proactive classroom management strategies, including praise statements and precorrections, and reactive strategies (i.e., use of reprimands), were gathered simultaneously during each observation. Prior studies have shown that these single 20-minute observations are significantly correlated with teacher self-reported classroom management self-efficacy and emotional exhaustion and are sensitive to change over time (Reinke et al., 2015). That is, teachers who received training to increase their use or proactive strategies had significantly higher BCIO-R scores compared to those who did not, controlling for baseline observations (Reinke et al., 2015; Reinke et al., 2018).

The MOOSES program utilizes second-by-second comparison of raters to determine reliability for each variable by determining a match between observers within a 5-second window. If a match was found, then an agreement for that variable was tallied. Variables that were not matched were tallied as disagreements. An agreement ratio was then reported for each variable (agreements divided by the sum of agreements plus disagreements). Ongoing reliability checks were conducted for between 32% to 42% of the observations across time points. The mean percentage agreement across time points on the BCIO-R was 92.3%, ranging from 90 to 95% for the four time points. Overall reliability of 80% is considered acceptable (Tapp, 2004).

Outcome Measures

Teacher report of child social behavior and academics. The *Teacher Observation of Classroom Adaptation-Checklist* (TOCA-C; Koth et al., 2009) is a 54-item measure of child behavior. It was completed by the classroom teachers for each child. Teachers rated each student at the beginning (September) and end (April/May) of the school year. They rated each child on the items referencing the past three weeks. The four subscales of the TOCA-C included in the present study were Disruptive Behaviors, Concentration Problems, Emotional Dysregulation, Internalizing, and Prosocial Behavior. Prior studies support the TOCA's internal consistency, consistent factor structure over time, predictive and current validity, and sensitivity to change across elementary and secondary school samples (Bradshaw et al., 2012; Koth et al., 2009). For the current study, the internal consistency (computed using Cronbach's alpha) for each subscale ranged from .77 to .96.

Teacher self-report of stress and coping. The burnout measure was derived from the Maslach Burnout Inventory (MBI; Maslach et al., 1996) and included four items from the emotional exhaustion subscale. Mean scores were computed based on these four items and were used in all analyses. While burnout is a multidimensional construct, as in previous studies of teacher stress, this study examined only the emotional exhaustion dimension of burnout. Emotional exhaustion is the primary experience of burnout most closely related to stress and coping and is defined by the experience of extended stress and low or ineffective coping over time (Pas et al., 2010). The internal consistency of the abbreviated scale for the current study was calculated using Cronbach's alpha; the alpha values in the study ranged from 0.82 to 0.95 (an average alpha of 0.91). Example items include, "Feel emotionally drained from work," and "Feel like at the end of the rope."

Teaching Coping Scale (Eddy et al., 2020). Teachers completed this measure at baseline and at the end of the school year. At baseline, the teachers were asked to rate their overall stress and coping using single-item measures of each construct. The stress question asked, "How stressful do you find being a teacher?", and the coping question asked, "How well are you coping with the stress of your job?" The questions stand-alone and no other instructions or details are given. The item scale ranged from 0 to 10 with 0 indicating "not stressful" and 10 indicating "very stressful" for the stress item and 0 indicating "not well" and 10 indicating "very well" for the coping item. A recent study found that these single-items predicted concurrent and prospective teacher burnout and self-efficacy and teacher practices (Eddy et al., 2020). Additionally, the items were used in a prior study to examine patterns of stress and coping in elementary school teachers and yielded strong profile fit that were associated with student academic and behavior outcomes as predicted (Herman et al., 2017).

Student self-report of depression. *Patient Health Questionnaire-8 Adolescent Version* (PHQ-8; Johnson et al., 2002). Students completed the PHQ-8 at baseline and again at the end of the school year; mean scores were computed. The PHQ-8 is a widely used measure of depressive symptoms that was adapted from the PHQ-9 (Kroenke et al., 2001). Prior studies have found the PHQ-8 and the PHQ-9 Adult and Adolescent versions demonstrate concurrent and criterion validity in community and clinical samples including with adolescents (Johnson et al., 2002). The 8 items map onto the diagnostic criteria for Major Depressive Disorder. The scale includes 4-point Likert responses (0-"not at all", 1-"several day", 2-"more than half the days" 3-"nearly every day"). An example item is, “Feeling down, depressed, irritable, or hopeless?” Internal consistencies for fall and spring of each study year ranged from 0.79 to 0.88.

Direct Behavior Rating (DBR)—Unhappy. DBR—Unhappy was modeled after the broader DBR scales (Chafouleas et al., 2009). Students rated their unhappiness on this single item scale (UN; Kilgus et al., 2019). This particular item was intended to serve as a broad and general indicator of student internalizing problems. Unhappy was defined as the expression of sadness, gloom, joylessness, or discontentment through words, body posture, tone of voice, facial expressions, or social cues. Examples included a limited range of facial expressions or animation, downward cast eyes and mouth, infrequent smiling or laughing, crying, inactivity, limited social participation, engagement in few pleasurable activities, low energy, recurrent expressions of worry or guilt, frequent physical complaints, pessimism, and negative self-statements.

Standardized academic achievement. Grade-Level Assessments (GLA). GLAs are assessed using the Missouri Assessment Program (MAP), which is a standardized, state-wide assessment administered to students in grades 3 through 8 in the spring of every school year. This criterion-referenced test was designed to measure student achievement toward state-level standards. Data included in the current study are from the end-of-year Mathematics and Communication Arts subtests of the MAP. Since 2014 the GLA assessments are online assessments administered by the district’s testing vendor. Scale scores produced for each student describes achievement on a continuum that spans 3rd to 8th grades. MAP scaled scores had acceptable Cronbach’s alpha coefficients. Specifically, reliability of the communication arts test was 0.87 for sixth grade, 0.90 for seventh grade, and .91 for eighth grade, and the mathematics test produced reliability coefficients of 0.88 for sixth grade, 0.90 for seventh grade, and 0.87 for the eighth grade versions of the test (Missouri Department of Elementary and Secondary Education, 2015).Within a content area MAP scores of adjacent grades can be compared.

Additionally, we administered subtests of the Stanford Achievement Test Tenth Edition (SAT-10; Harcourt, 2004) pre, post, and in the spring of the following year. The SAT-10 is a widely used group-administered standardized measure of academic achievement developed around national and state curriculum standards as well as those trends promoted by national professional educational groups. It is designed to estimate academic achievement in reading, math, language arts, and science. Extensive research documents the reliability and construct validity of the SAT-10 (Harcourt, 2004). Subtest coefficient alphas all exceed .80. We used two subtests, the Reading Comprehension subtests for students in reading/English classes and the Problem Solving subtest for students in math classes. Assessment occurred post intervention in April and May of the same school year.

Student demographics. Free and reduced lunch status (FRL), race, sex, age, grade, and special education status were obtained from the school district for all participating students. Students were coded as 1 if they received FRL and 0 if not. Student sex was coded as 1 for female and 0 for male. Students receiving special education were coded as 1 and if not 0. For the purposes of this study student race was coded as Black, White, or Other Race.

Analytic Approach

For the data analysis, we used multiple imputation for handling missing data (Schafer & Olsen, 1998). We checked covariate balance by calculating the effect sizes of the covariates among four treatment-by-moderator (event) subgroups. We then used hierarchical linear models (HLM) to account for nested data structure (e.g., students nested within teachers, repeated measured nested within teachers) by controlling for the baseline covariates for the analysis of the teacher and student outcomes.

Missing data. Missing data occurred primarily on the outcome measures. The missing rates for the pretests of four social and behavioral outcome measures is 0.5% while the missing rates for the posttests of four social and behavioral outcome measures is 14.2% in the overall sample; the vast majority of this missing data was the result of students moving out of the school district during the year. The differential missing rates between the treatment and control groups are -0.7% for the pretest and 2.7% for the posttest. Based on the What Works Clearinghouse (WWC; 2014) attrition standard, the combination of an overall attrition rate of 14.2% and a differential attrition rate of 2.7% would result in low levels of potential bias (i.e., greater than 0.05 of standard deviation) even under the more conservative assumptions. Hence, the results from the analysis of the students who have complete posttests will have good internal validity. The literature also showed that when the outcome is included in the imputation model, there are very small differences between models that impute the outcome compared with those that do not (Kontopantelis et al., 2017). The final analytic samples in Cohorts 1-3 included 9 schools 85 teachers and 1069 students for the analyses of social and behavioral outcomes; among 85 teachers 39 teachers in math class (497 students for Problem Solving and 502 for MAP Math) and 46 teachers in reading class (552 students for Reading Comprehension and 563 students for MAP English) for the analyses of academic achievement outcomes.

The maximum overall data missing rate and differential missing rate between the treatment and control group in the final analytic samples for the analysis of social and behavioral outcomes were 0.6% and -0.8%, respectively. The maximum overall data missing rate and differential missing rate between the treatment and control group in the final analytic samples for the analysis of academic outcomes were 11.2% and 3.0%, respectively. Multiple imputation using a Markov chain Monte Carlo (MCMC) method in SAS PROC MI was used to impute missing data by including posttest, pretest, and other covariates. We imputed five times for the final analytic samples for the analysis of social and behavioral outcomes and 30 times for academic outcomes based on the missing rates (Schafer & Olsen, 1998).

Analysis of teacher implementation. First, to evaluate whether teacher implementation of proactive classroom management skills increased following receipt of the CHAMPS intervention, we conducted longitudinal analysis. We fit a linear growth curve model using two-level hierarchical linear modeling (HLM) using SAS PROC MIXED. The repeated measures (level 1) are nested within teachers (level 2). We controlled for the baseline pretest in evaluating the treatment effects on teacher implementation of proactive classroom management skills. We also calculated the mean rate of praise, precorrections, and reprimands observed at each time point to demonstrate any changes in the base rate of the teacher behaviors.

Analysis of main and moderator effects on student outcomes. For each of the five imputed datasets, two-level hierarchical linear models (HLM), in which students (level 1) are nested within teachers (level 2), were conducted using SAS PROC MIXED to examine the overall treatment effects student behavior and academic outcomes. Each student's pretest and demographic information were included at level 1, and the treatment variable was at level 2. SAS PROC MIANALYZE was used to combine the results from the analyses of five datasets. The full statistical model is below:

Level 1 (student): $Y_{ij} = \beta_{0j} + \beta_{1j}Black_{ij} + \sum_{q=2}^{Q}\beta_{qj}\boldsymbol{X}_{qj} + r_{ij}, r_{ij} \sim N(0, \sigma^2)$

Level 2 (class): $\beta_{0j} = \gamma_{00} + \gamma_{01}Treatment_j + \gamma_{02}Event_j + \gamma_{03}Treatment_j * Event_j + u_{0j}, u_{0j} \sim N(0, \tau)$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Treatment_j + \gamma_{12}Event_j + \gamma_{13}Treatment_j * Event_j$$
$$\beta_{qj} = \sum_{q=2}^{Q}\gamma_q$$

Where $Black_{ij}$ is binary variable that represents whether a student *i* in class *j* is a black student or not ($Black$ = 1 for black students, = 0 otherwise), $\boldsymbol{X}_{qj}$represent student-level covariates, which include pretest, age at pretest, gender, FRL, special education status, grade level, and cohort year in the study. $Treatment_j$ is a binary variable indicating treatment condition (Condition = 0 for control group and Condition = 1 for treatment group). $Event_j$ is a binary variable that represents the Mike Brown event ($Event$ = 0 for before the Mike Brown event, and = 1 for after the Mike Brown event). The parameter, γ_{01}, estimates the treatment effect for non-Black students before the Mike Brown event. The parameter, γ_{02}, estimates the effect of the Mike Brown event for the non-Black students in the control group. The parameter, γ_{03}, estimates the moderated effect of the Mike Brown event on the intervention for the non-Black students. The parameter, γ_{10}, estimates the Black-White gap for students in the control group before the Mike Brown event. The parameter, γ_{11}, estimates the additional (moderated) treatment for the Black students (or the additional B-W gap for students in the treatment group) before the event. The parameter, γ_{12}, estimates the additional effect of the Mike Brown event for the Black students (or the additional B-W gap for White students after the Mike Brown event). The parameter, γ_{13}, estimates the additional treatment effect for the Black students after the Mike Brown event (or the additional B-W gap for Black students after the Mike Brown event). σ^2 and τ are variance components for level 1 and level 2 residuals conditional on these variables. Note that we tested the full model and we also tested simplified models by removing non-significant interaction terms. We also tested the baseline measures as moderators by using the treatment condition to predict the coefficients of the baseline measures (model omitted for simplicity).

Results

Descriptive Statistics and Covariate Balance Checking

Table 1 provides descriptive statistics and covariate balance checking for the analytic sample of social behavioral outcomes at baseline. The maximum effect sizes among four treatment-by-event groups are also provided in Table 1. Most baseline measures were balanced among four groups. We included all the covariates in the HLM to reduce bias.

Event Effects on Teacher Adherence to CHAMPS

A two-level HLM of STOIC ratings at the end of the school year, adjusting for baseline climate scores, revealed that there was no significant intervention effect for non-Black teachers before the event (*d* = 0.02, *p* = 0.87), no significant event effect for the non-Black teachers (RQ 1), and no moderated treatment effect for the event (*d* = 0.22, *p* = 0.25; RQ 2). However, there was a significant event effect for Black teachers (*d* = 0.64, *p* < 0.0001; RQ 3), significant moderated treatment effects for Black teachers before the event (*d* = 0.54, *p* = 0.0031; RQ 4), and significant moderated treatment effects for Black teachers after the event (*d* = -0.66, *p* = 0.0048; RQ 5) (Table 2). We further present the results regarding treatment effect sizes on different groups and the Black-White gap (Black effect size minus non-Black effect size) under different conditions in Figure 1. For example, the treatment effect for Black teachers before the event was significant (*d* = 0.95, *p* < 0.0001) while the treatment effects for other groups were not significant; the Black-White gap in the control group before the event was significant (*d* = -0.54, *p* = 0.0061), the Black-White gap in the control group after the event was also significant (*d* = 0.53, *p* = 0.0011), while the Black-White gaps under other conditions were not significant. These results indicate that the CHAMPS intervention had a stronger positive effect on Black teachers across time periods, but this effect was reduced after the event. Additionally, in the Control group, the Black-White gap in classroom management skills favored White teachers before the event, but favored Black teachers after the event. The intervention ultimately made the Black-White gap non-significant.

Teacher Implementation of Proactive Classroom Management

To evaluate whether teachers receiving CHAMPS demonstrated an increase in their implementation of proactive strategies in comparison to control teachers, a two-level HLM was conducted on BCIO-R positive-to-negative ratios (see Table S-1) controlling for the baseline positive-negative ratio. Analyses on teacher implementation of positive-negative strategies revealed that there was no significant intervention effect for non-Black teachers before the event (*b* = 7.75, *p* = 0.2209), no significant event effect for the non-Black teachers (RQ 1), no moderated treatment effect for the event (*d* = 11.28, *p* = 0.3073; RQ 2), and no significant moderated treatment effect for Black teachers before the event (*d* = 13.01, *p* = 0.2823; RQ 4). However, there was a significant event effect for Black teachers (*d* = 39.13, *p* = 0.0001; RQ 3), and significant moderated treatment effect for Black teachers after the event (*d* = -68.08, *p* = 0.0002; RQ 5). We further present the results regarding treatment effect sizes on different groups and Black-White gap under different conditions in Figure S-1. For example, the treatment effect for non-Black teachers after the event is significant (*d* = 0.63, *p* = 0.0319), the treatment effect for Black teachers before the event is significant (*d* = 0.69, *p* = 0.0450), and the treatment effect for Black teachers after the event is significant (*d* = -1.20, *p* < 0.0001) while the treatment effect for non-Black before the event is not significant (*d* = 0.26, *p* = 0.2209); the Black-White gap in the control group after the event is significant (*d* = 1.00, *p* < 0.0001), the Black-White gap in the treatment group after the event is also significant (*d* = -0.83, *p* = 0.0002), while the Black-White gaps under other conditions are not significant.

These results indicate that CHAMPS had a significant and moderate benefit for non-Black teachers positive-to-negative ratio, but only after the event. On the other hand, CHAMPS significantly and moderately improved Black teachers positive-to-negative ratio before the event only but their ratio worsened (became more negative) after the event. The latter reduction of Black teacher positive-to-negative ratio after the event represented a large effect. While the Black-White gap in these ratios was not significant in either condition before the event, Black teachers in the control condition significantly outperformed their non-Black counterparts after the event whereas Black teachers in the treatment condition significantly underperformed their non-Black counterparts after the event. Both of these differences represented large effects.

Differential Effects on Student Social Behavior

The baseline distress measures did not moderate event-related treatment effects on any teacher or student outcomes (RQ 6). The Mike Brown event or being Black did not have any significant moderation effect on the teacher-reported concentration problems, teacher-reported disruptive behavior problems, or teacher-reported emotional dysregulation (RQs 2 & 4). There were some moderation effects on prosocial behavior (Table 3). There was no significant effect of the Mike Brown event on White students in the control group on prosocial behavior (d = -0.04, p = .6517; RQ 1), and there was no significant intervention effect on Black students before the event and non-Black students before and after the event (d = 0.04, p = .5133). However, the differential impact of the event on prosocial behavior between Black and non-Black students was significant (d = -0.32, p = .0144; RQ 4), and the additional treatment effect on Black students after the event was significant (d = 0.37, p = .0194; RQ 5). We further present the results regarding treatment effect sizes on different groups and Black-White gap under different conditions in Figure 2. For example, the treatment effect for Black students after the event is significant (d = 0.40, p = 0.0065) while the treatment effect for non-Black students before the event is not significant (d = 0.04, p = 0.5133); the Black-White gap in the control group after the event is significant (d = -0.34, p = 0.0024) while the Black-White gaps under other conditions are not significant. These results indicate that CHAMPS had a small significant effect on Black students after the Michael Brown event. In the control group, Black students experienced a small and significant worsening of prosocial skills after the event compared to non-Black students..

Differential Effects on Student Academic Outcomes

There were no significant moderation effects on the MAP Math or SAT-10 Reading Comprehension scores. However, the event had a significant additional effect on Black students (d = -17.58, p = 0.0142; RQ 3) on SAT-10 Problem Solving scales (Tables S-2). Figure 3 illustrates the

Black-White gaps on Problem Solving scales before the Brown event (*d* = -0.07, *p* = 0.5258) and after the event (*d* = -0.59, *p* < 0.0001). The intervention had a significant effect on non-Black students before and after the event and Black students before the event on MAP Communication Arts (*d* = 0.28, *p* = 0.0002) and intervention had a significant additional negative effect on Black students after the event (*d* = -0.39, *p* = 0.0005; RQ 5; Table S-3), which results in non-significant treatment effect for Black students after the event (*d* = -0.19, *p* = 0.1108) (Figure S-2).

Discussion

The findings suggest that a traumatic historical event, the shooting of Michael Brown, differentially affected teachers and students in a RCT evaluation of a classroom management program conducted before and after the shooting in a nearby school district. Specifically, Black teachers benefitted more from the CHAMPS intervention compared to non-Black teachers as evidenced by their independently observed classroom management skills and praise-to-reprimand ratios; however, these effects were minimized or disappeared after the Michael Brown event. In contrast, the intervention significantly increased non-Black teacher positive-to-negative ratios, but only after the event. Additionally, the Black-White achievement gap favoring non-Black students significantly increased after the event. In particular, although the intervention had a significant and small benefit for both non-Black and Black students' English achievement scores before the Michael Brown event, the treatment effect for Black students was no longer significant after the event.

The study also revealed some differences in the teacher control group before and after the event. Before the event, Black control group teachers had lower levels of positive interactions and effective classroom management practices compared to their non-Black peers. After the event, however, Black teachers in the control group had significantly higher levels of positive interactions and effective classroom management practices and, in turn, the Black-White control group gap flipped to favor Black teachers after the event. Combined with the finding that the event was associated with a worsening of Black teacher performance in the treatment group suggests that the event differentially impacted teachers based on their race and treatment status. We can speculate that Black teachers in the intervention condition were adversely impacted by the event given the shooting involved a Black high school student and ongoing stressors related police brutality against Black citizens. It is possible they perceived the intervention and coaching to be overly burdensome in the context of the stressors they experienced after the event. It is also worth noting that their intervention coach was a White female so this racial disparity may have intersected with the event's effects to make Black teachers less interested in or able to implement the intervention. Whereas in the Control group that did not have the intervention or access to a coach, findings suggest that Black teachers improved their interactions and classroom management. It may be that Black teachers in the absence of an intervention developed a stronger sense of responsibility for their mostly Black student population and provided more positive and structured learning contexts to support them during the time of stress and community trauma.

In addition to the negative effect of the event on Black student achievement, findings also revealed a counterintuitive student effect. The intervention had a significantly stronger effect on the teacher-rated prosocial skills of Black students after the event compared to other students before and after the event. Notably, Black students in the control condition also had a significant deterioration in their prosocial skills after the event compared to other groups. Thus, it appears the intervention served to reduce the harmful effects of the Michael Brown event on student behaviors that were observed in the absence of intervention. Given the higher quality of classroom management delivered by CHAMPS Black teachers before and after the event as well as the higher levels of positive-to-negative ratio delivered by CHAMPS non-Black teachers after the event, these structured, positive environments may have mitigated the deterioration of student prosocial behaviors observed in Black students in the control condition.

While the findings are interesting and important, one limitation is that although the treatment was randomly assigned, the teachers and students were not randomly assigned to different event groups. Our covariate balance check did not reveal large differences on the measured baseline covariates among four treatment-by-moderator subgroups and we controlled all the covariates in our analysis; however, there may be potential hidden bias due to omitted variables confounding with the event.

Implications

The findings speak to the power of contexts in influence educational outcomes for students. Our prior study revealed the benefit of the program on average for all teachers and students (AUTHOR, 2020. Here we found evidence that the intervention was particularly helpful for Black teachers, especially in the absence of the historical event. The reduced effect on Black teachers after the event and the subsequent improvement in control Black teachers suggests that the effect size of the intervention on Black teachers may be underestimated in the context of an intervention training not proximal to a traumatic community event. The negative effects for Black youth achievement after the event suggests the damage that ongoing community turmoil, in this case, specific to police brutality against Black citizens and the expansion of Black-White achievement gaps in these contexts. Thus, school-based interventions alone will not likely reduce these gaps in the context of large sociocontextual challenges presented to Black youth identity and safety. As long as grand social inequities persist outside of school, including disproportional police brutality experienced by Black citizens, these circumstances may undermine the impact of even the most effective educational interventions.

This has implications for the training, recruitment, and retention of Black teachers. Much has been written about the shortage of Black teachers and the need to increase Black representation in the teaching ranks (Rogers-Ard et al., 2013). CHAMPS and similar interventions focused on positive classroom management skills holds promise as a tool to make this happen, particularly in the absence of a traumatic historical event.

The findings support the importance of contextualizing educational research, including experimental designs. In the present study, we found evidence to show that a proximal traumatic event had a strong influence over the effect size of an intervention that varied based on the demographic characteristics of participants and their treatment status. Although this event was a prominent contextual feature, it is likely that other less momentous conditions may influence intervention dose and quality that should be considered in all human subject studies (Kaplan et al., 2020) and that it is important for investigators to carefully examine and document in their reports.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review, 42(1),* 76-98.

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(5),* e1136–e1145.

Brewin, C. R., Andrews, B., & Valentine, J. D. (2000). Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. *Journal of Consulting and Clinical Psychology, 68(5),* 748-766.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54(4),* 297.

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009). Direct behavior rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention, 34,* 195–200.

Eddy, C. L., Herman, K. C., & Reinke, W. M. (2019). Single-item teacher stress and coping measures: Concurrent and predictive validity and sensitivity to change. *Journal of school Psychology, 76,* 17-32.

Harcourt Assessment, Inc. (2004). *Stanford Achievement Test series—Tenth edition technical data report.* San Antonio, TX: Author.

Herman, K. C., Hickmon-Rosa, J. E., & Reinke, W. M. (2018). Empirically derived profiles of teacher stress, burnout, self-efficacy, and coping and associated student outcomes. *Journal of Positive Behavior Interventions, 20(2),* 90-100.

Herman, K. C., Reinke, W. M., & Dong, N. (2020). *Can effective classroom behavior management increase student schievement in middle school? Findings from a group randomized trial.* Manuscript under review.

Johnson, J. G., Harris, E. S., Spitzer, R. L., & Williams, J. B. (2002). The patient health questionnaire for adolescents: validation of an instrument for the assessment of mental disorders among adolescent primary care patients. *Journal of Adolescent Health, 30(3),* 196-204.

Kaplan, A., Cromley, J., Perez, T., Dai, T., Mara, K., & Balsai, M. (2020). The role of context in educational RCT findings: A call to redefine "Evidence-Based Practice". *Educational Researcher, 49(4),* 285-288.

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., ... & Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence, 95,* S5-S28.

Kilgus, S. P., Van Wie, M. P., Sinclair, J. S., Riley-Tillman, T. C., & Herman, K. C. (2019). Developing a direct rating behavior scale for depression in middle school students. *School Psychology, 34(1),* 86.

Kochel, T. R. (2015). Assessing the initial impact of the Michael Brown shooting and police and public responses to it on St Louis County residents' views about police (Report). Department of Criminology and Social Justice: Southern Illinois University Carbondale.

Kontopantelis, E., White, I. R., Sperrin, M., & Buchan, I. (2017). Outcome-sensitive multiple imputation: a simulation study. *BMC Medical Research Methodology, 17(1)*, 2.

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.

Kroenke, K., Spitzer, R.L., & Williams, J.B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-13.

Maslach, C., Jackson, S.E., & Leiter, M.P. (1996). *Maslach Burnout Inventory manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Missouri Department of Elementary and Secondary Education (2015). *Missouri Assessment of Progress test.* Jefferson City: DESE.

National Institute of Child Health and Human Development, Early Child Care Research Network (NICHD ECCRN) (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *Elementary School Journal, 105*, 305 – 323. https://doi.org/10.1086/428746

Pas, E. T., Bradshaw, C. P., & Mitchell, M. M. (2011). Examining the validity of office discipline referrals as an indicator of student behavior problems. *Psychology in the Schools, 48*, 541-555.

Pas, E. T., Bradshaw, C. P., Hershfeldt, P. A., & Leaf, P. J. (2010). A multilevel exploration of the influence of teacher efficacy and burnout on response to student problem behavior and school-based service use. *School Psychology Quarterly, 25(1),* 13.

Petras, H., Chilcoat, H., Leaf, P., Ialongo, N., & Kellam, S. (2004). The utility of teacher ratings of aggression during the elementary school years in identifying later violence in adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry, 1*, 88–96.

Pianta, R. C., Hamre, B. K., Hayes, N., Mintz, S., & LaParo, K. M. (2008). Classroom assessment scoring system - Secondary (CLASS-S). Charlottesville, VA: University of Virginia.

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.

Reinke, W. M., Stormont, M., Herman, K. C., Wachsmuth, S., & Newcomer, L. (2015). The Brief Classroom Interaction Observation–Revised: An observation system to inform and increase teacher use of universal classroom management practices. *Journal of Positive Behavior Interventions, 17(3*), 159-169.

Rogers-Ard, R., Knaus, C. B., Epstein, K., & Mayfield, K. (2013). Racial diversity sounds nice; system transformation? Not so much: Developing urban teachers of color. *Urban Education, 48(3),* 451–479.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33(4),* 545-571.

Sprick, R. (2013). *STOIC observation*. Eugene, OR: Safe & Civil Schools.

Sprick, R., Garrison, M., & Howard, L. (2009). CHAMPS: *A proactive and positive approach to classroom management* (2nd edition). Eugene, OR: Pacific Northwest Publishing.

Strickler, R., & Lawson, E. (2020). Racial conservatism, self-monitoring, and perceptions of police violence. *Politics, Groups, and Identities*, 1-22 (Online First).

Tapp, J. (2004). *MOOSES* (Multi-Option Observation System for Experimental Studies). http://kc.vanderbilt.edu/mooses/mooses.html.

Vogt, D. S., King, D. W., & King, L. A. (2007). Risk pathways for PTSD (pp. 99-115). In M.J. Friedman, T. M. Keane, and P. A. Resick (Eds.), *Handbook of PTSD: Science and practice*. New York: Guilford.

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Washington DC: Institute of Education Sciences.

Table 1

Covariate balance checking among four treatment-by-event subgroups for the analytic sample of social behavioral outcomes at baseline

Treatment Status	Control		Control		Treatment		Treatment		Maximum ES
Brown Event	Before		After		Before		After		
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Age	12.80	0.88	12.31	0.87	12.47	0.84	12.35	0.86	0.56
Female	0.46	0.50	0.48	0.50	0.54	0.50	0.46	0.50	0.16
Lunch Status	0.66	0.48	0.62	0.49	0.62	0.49	0.55	0.50	0.21
Special Education	0.07	0.26	0.12	0.33	0.06	0.24	0.14	0.35	0.28
White	0.14	0.35	0.41	0.49	0.14	0.35	0.40	0.49	0.68
African American	0.82	0.38	0.55	0.50	0.81	0.39	0.55	0.50	0.66
Other race	0.04	0.19	0.05	0.21	0.05	0.21	0.05	0.22	0.06
Grade 6	0.29	0.45	0.54	0.50	0.46	0.50	0.56	0.50	0.56
Grade 7	0.37	0.48	0.28	0.45	0.36	0.48	0.23	0.42	0.30
Grade 8	0.34	0.48	0.18	0.39	0.18	0.38	0.21	0.41	0.39
TOCA - concentration problems	3.03	1.28	2.70	1.23	2.83	1.18	3.04	1.38	0.28
TOCA - disruptive behavior	1.88	0.77	1.77	0.71	1.71	0.64	1.85	0.83	0.23
TOCA - prosocial behavior	4.50	0.96	4.68	0.88	4.58	0.91	4.48	0.99	0.22
TOCA - emotion regulation	2.34	1.09	2.23	1.04	2.24	0.96	2.38	0.99	0.15
TOCA -internalizing	1.72	0.73	1.86	0.70	1.76	0.71	1.76	0.83	0.19
PHQ8 - depression	0.11	5.20	-0.25	5.52	-0.21	4.63	0.51	5.19	0.15
DBR - unhappy	0.03	1.99	-0.18	1.93	-0.14	1.77	-0.04	2.05	0.11
N	381		170		395		123		
J	31		12		30		12		

Table 2

HLM Results for 2-Level Model Examining the Effects of CHAMPS on STOIC Classroom Management

Variable	*b*	SE	*p*-value
Intercept	2.23	0.28	0.0000
Time	-0.04	0.03	0.2812
CLASS-S climate	0.25	0.06	0.0000
Teacher coping	0.06	0.02	0.0007
Black	-0.32	0.12	0.0061
Treatment	0.02	0.12	0.8714
Event	-0.14	0.14	0.3037
Treatment *Event	0.22	0.19	0.2450
Treatment * Black	0.54	0.18	0.0031
Event * Black	0.64	0.15	0.0000
Treatment * Event * Black	-0.66	0.23	0.0048

Table 3

HLM Results for 3-Level Model Examining the Effects of CHAMPS on Student Prosocial

Variable	*b*	SE	*p*-value
Intercept	1.21	0.67	0.0702
Age	0.00	0.06	0.9416
Female	0.11	0.04	0.0095
Lunch Status	-0.04	0.05	0.4065
Special Education	-0.06	0.09	0.4774
African American	-0.02	0.08	0.8076
Other Race	0.19	0.09	0.0321
Grade 7	0.08	0.10	0.4172
Grade 8	-0.03	0.14	0.8423
DBR - Unhappy	-0.03	0.01	0.0047
TOCA- prosocial	0.71	0.03	0.0000
Event	-0.04	0.08	0.6517
Treatment	0.04	0.06	0.5133
Event * Black	-0.32	0.13	0.0144
Treatment * Event * Black	0.37	0.16	0.0194

Note. TOCA = *Teacher Observation of Classroom Adaptation-Checklist.* DBR = *Direct Behavior Rating.*

Figure 1

Treatment Effects and Black-White Gaps before and after the Event on Teacher STOIC Classroom Management

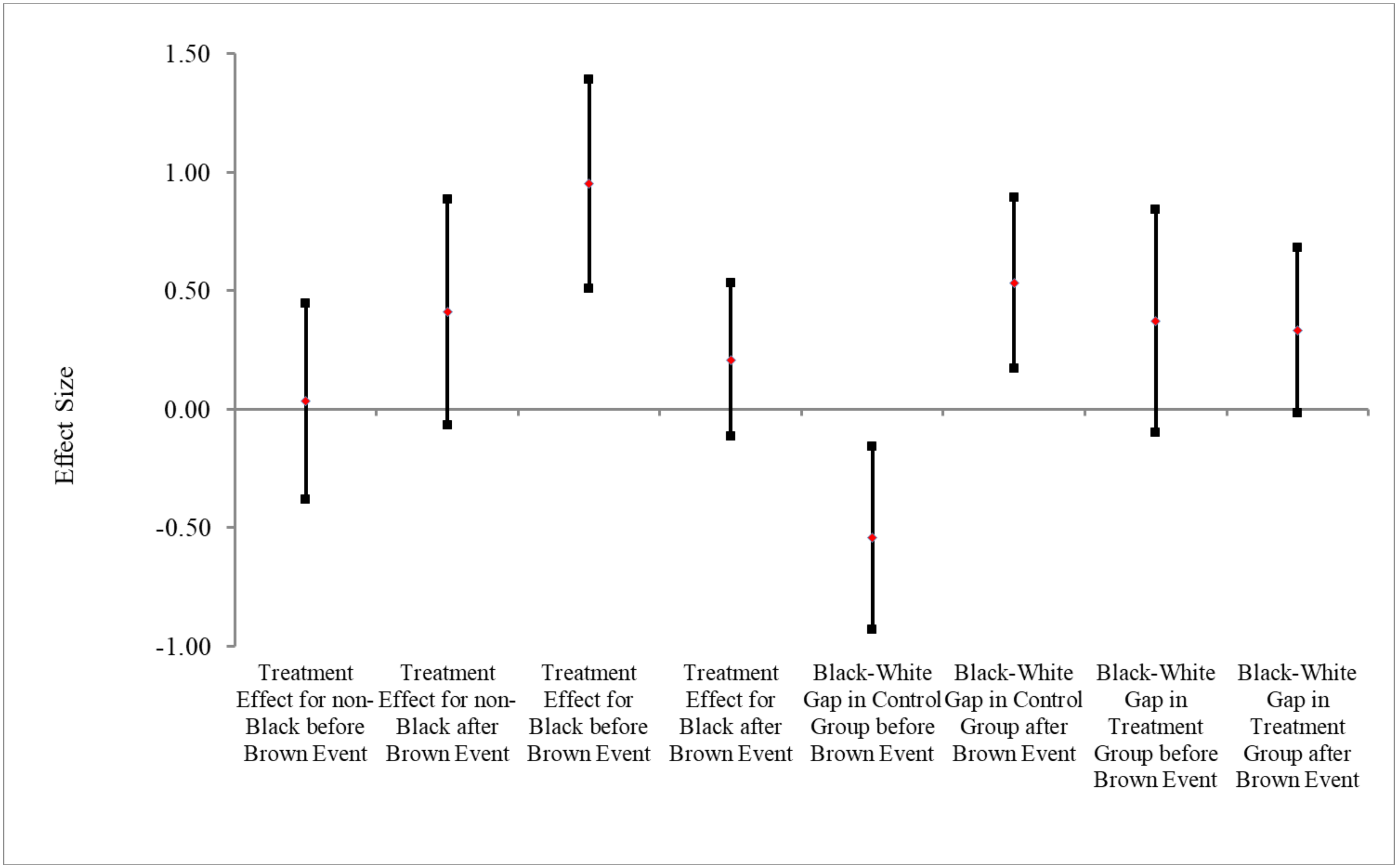


Figure 2

Treatment Effects and Black-White Gaps before and after the Event on Student TOCA-C Prosocial Behaviors

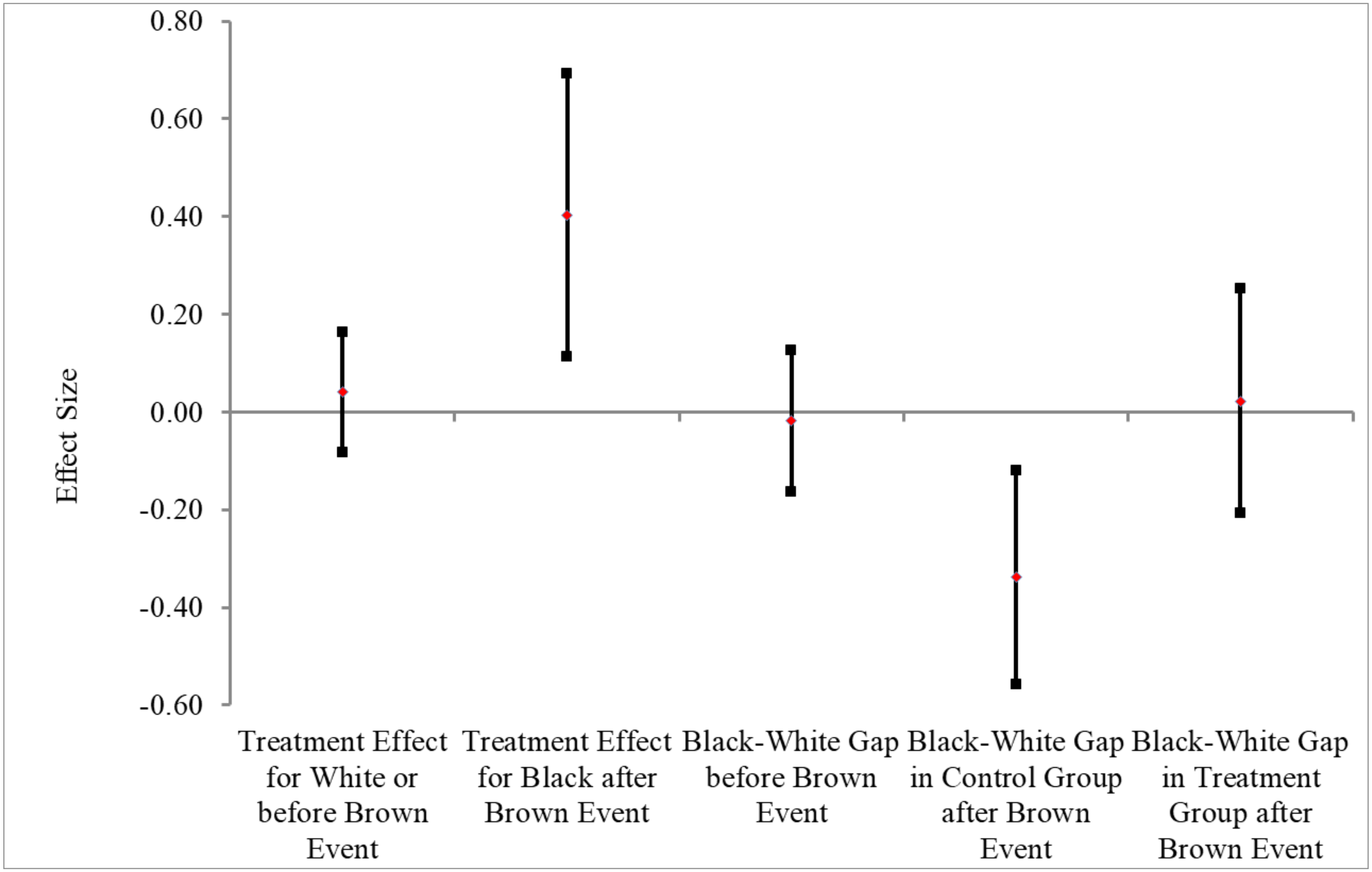
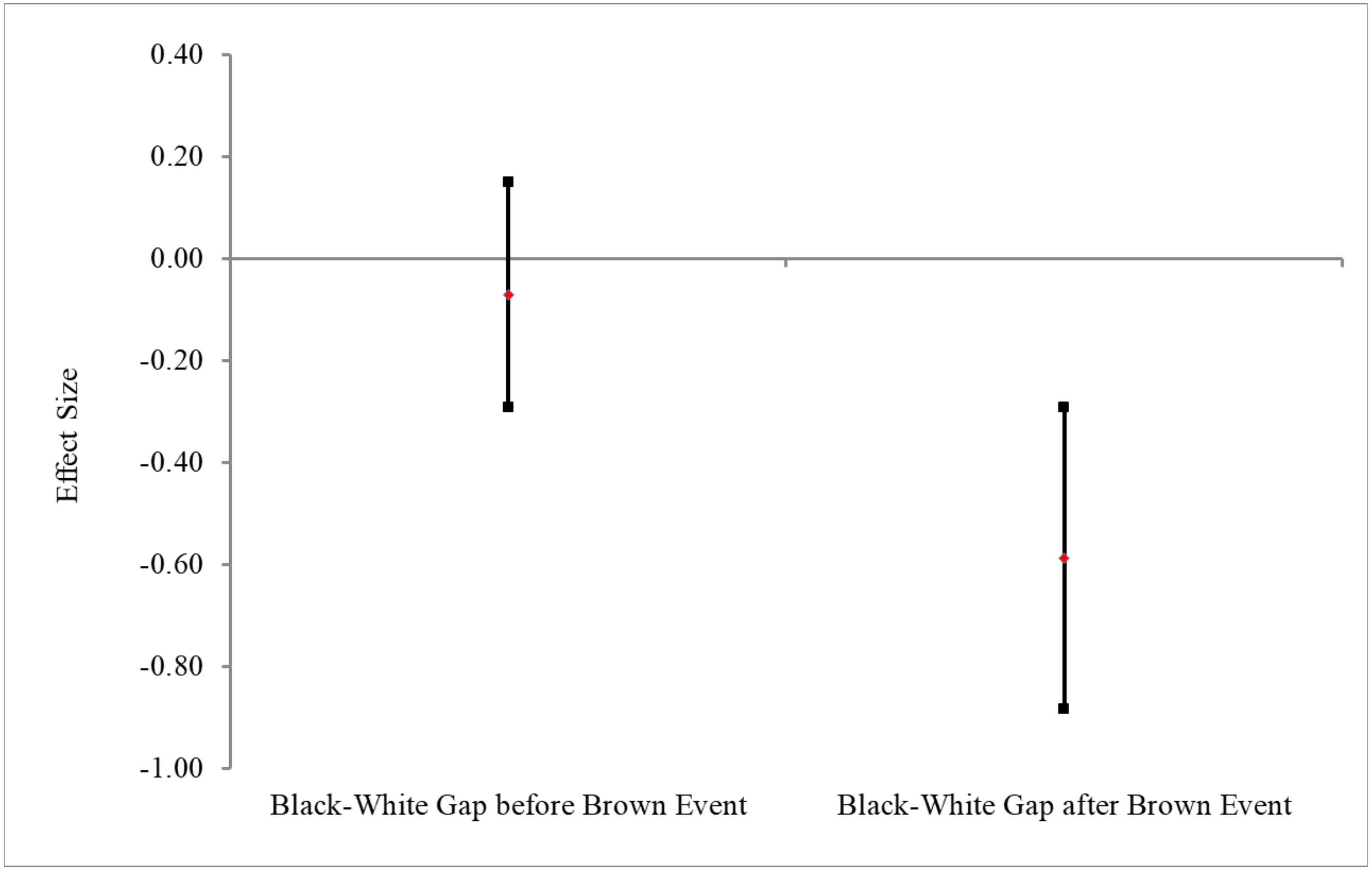


Figure 3

Black-White Gaps before and after the Event on Student SAT-10 Problem Solving


Effect Size
Black-White Gap before Brown Event
Black-White Gap after Brown Event