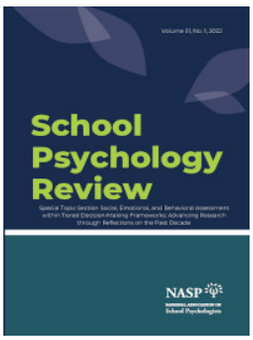


School Psychology Review

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uspr20

Accounting for Traumatic Historical Events in Educational Randomized Controlled Trials

Keith C. Herman, Nianbo Dong, Wendy M. Reinke & Catherine P. Bradshaw

To cite this article: Keith C. Herman, Nianbo Dong, Wendy M. Reinke & Catherine P. Bradshaw (2022): Accounting for Traumatic Historical Events in Educational Randomized Controlled Trials, School Psychology Review, DOI: 10.1080/2372966X.2021.2024768

To link to this article: https://doi.org/10.1080/2372966X.2021.2024768

SCHOOL PSYCHOLOGY REVIEW
https://doi.org/10.1080/2372966X.2021.2024768

Accounting for Traumatic Historical Events in Educational Randomized Controlled Trials

Keith C. Herman[a], Nianbo Dong[b], Wendy M. Reinke[a], and Catherine P. Bradshaw[c]

[a]University of Missouri; [b]University of North Carolina at Chapel Hill; [c]University of Virginia

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 was conducted in a nearby district before and after the shooting of Michael Brown in Ferguson, Missouri ($N = 102$ teachers and 1,450 students). The findings suggest that the event differentially affected teacher and student response within and across conditions. Black teachers benefited more from the intervention 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 benefited the academic achievement of Black and White students before the event, the opportunity gap widened after the event. Implications for the design, analysis, and reporting of findings from randomized controlled trials are discussed.

IMPACT STATEMENT

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. Moreover, findings suggest that the Black–White opportunity gap will not be reduced by schools alone.

ARTICLE HISTORY
Received June 28, 2021
Accepted December 22, 2021

KEYWORDS
history, opportunity gap, classroom management, academic achievement

ASSOCIATE EDITOR
Tyler Renshaw

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 postintervention 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 posttest 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.

The COVID-19 pandemic and global protests about police brutality that occurred during the spring of 2020 are two prominent historical events that may impact the dose and effect size of experimental studies in educational settings. These events have been experienced in one way or another by nearly every adult and child in the U.S. It is not known how these experiences will affect the findings derived from randomized control trials (RCTs) conducted during this period.

One way to examine how historical events may affect studies is to consider how traumatic the event might be for participants. Several elements have been identified that

CONTACT Keith C. Herman hermanke@missouri.edu Missouri Prevention Science Institute, University of Missouri, Columbia, MO, USA.
Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/2372966X.2021.2024768
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

 DOI:10.1080/2372966X.2021.2024768

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). For instance, Sharkey et al. (2014) found that proximity to violent crime incidents significantly reduced student performance on language achievement tests. 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. Notably, in the Sharkey et al. study, Black student performance particularly suffered in the aftermath of violent crime. Exposure to crime lowered Black students' likelihood to pass a language exam by 3 percentage points, an effect equal to 18 percent of the Black–White gap in passing rates. As another example, 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 preexisting 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 was 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 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. The RCT took place in one of the nearby school districts; all of the schools in the study were within a 10-mile radius of where the shooting occurred.

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). In particular, exposure to proximal police brutality events has a negative effect on the mental health of Black citizens. In one study with over 100,000 Black respondents, increasing numbers of police killings of unarmed Black citizens in a given state during the 3 month period preceding the survey was associated with incremental worsening of mental health for Black respondents in those states (Bor et al., 2018). Based on the effect sizes in their study, Bor and colleagues concluded that "the population mental health burden from police killings among Black Americans is nearly as large as the mental health burden associated with diabetes (p. 308)."

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). CHAMPS is a widely disseminated training program that focuses on supporting teacher use of effective classroom management practices including creating structured and predictable classroom routines, establishing clear expectations, monitoring student behavior, interacting positively with students, and correctly misbehaviors fluently and without emotion. Training takes place over several days and teachers are supported by a coach to implement better practices in their classrooms. The study design to evaluate the effects of CHAMPS 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.

The primary outcome analyses revealed that the CHAMPS training program significantly improved teacher proactive practices and student concentration problems, time-on-task, work completion, problem solving, and communication skills (Herman et al., 2020). Given recent events, we were motivated to examine any differential treatment effects on the third cohort of teachers and students that experienced the aftermath of the 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 (Saleem et al., 2020). 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 (Vogt et al., 2007).

Theoretical Framework

A recent theory, Developmental and Ecological Model of Youth Racial Trauma (DEMYth-RT), provides a

comprehensive approach to understand how traumatic experiences related to racial identity may influence youth response to interventions (Saleem et al., 2020). Race is a socially constructed categorization of people based on shared physical characteristics and customs that has been used to create social hierarchies, and racism is "a system of beliefs, practices, and policies that operate to advantage those at the top of the racial hierarchy" (p. 886; Haeny et al., 2021). DEMYth-RT examines the lasting and often ignored effects of historical, interpersonal, and vicarious encounters of racism and racial discrimination experienced by youth of color. It provides a developmental and ecological lens for understanding how racial trauma can influence individuals and the systems that surround them and can be passed across generations. The immediate and persisting effects of racial trauma on youth include psychological symptoms such as negative mood, avoidance, hyperarousal, and intrusive thoughts. During early adolescence, racial trauma can interfere with identity development and may lead to preoccupation with personal and family safety which in turn leads to distractibility and lower school performance (Saleem et al., 2020).

Many studies have examined the consequences of proximal or interpersonal consequences of racial trauma on youth mental health and well-being. In a review of the literature, Priest et al. (2013) identified 121 studies that focused on racial discrimination experiences of youth of color; 76% of the outcomes reported across these studies found statistically significant associations between racial discrimination and adverse mental health outcomes. However, the authors noted that the literature was limited by the lack of longitudinal studies and poor psychometric properties of many scales that measure discrimination experiences. Additionally, few studies have examined the relations between broader structural/cultural racism on youth outcomes or how such racism may moderate intervention response for youth of color (Price et al., in press; Saleem et al., 2020).

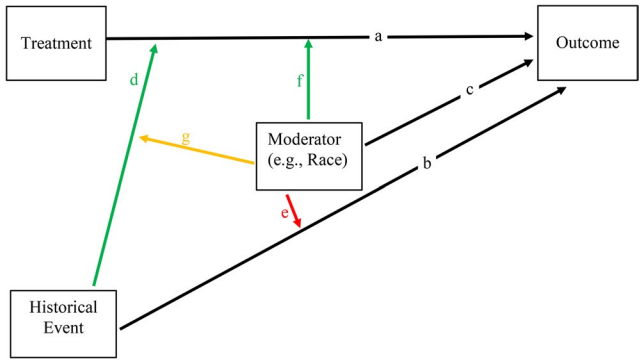
Of relevance, two studies with adults found evidence that interventions to prevent human immunodeficiency virus (HIV) exposure were significantly less effective in contexts characterized by high levels of anti-Black racism and sexism (Price et al., in press; Reid et al., 2014). For instance, Reid et al. (2014) examined 70 independent studies with 50% or greater Black participant samples and 99 unique interventions to increase condom use. Both negative White attitudes toward Black citizens and higher segregation levels of the communities where these studies occurred predicted lower intervention effect sizes. The authors concluded that structural stigma and discrimination reduced the efficacy of the intervention for Black participants.

We could identify only one parallel study with Black youth. Price et al. (in press) recently examined the effectiveness of psychotherapy for Black youth in settings that varied by anti-Black racism. The authors conducted a spatial meta-analysis of youth psychotherapy literature from 1963 to 2017 with 194 studies, 14,081 participants, across 34 states. They defined anti-Black cultural racism as a composite index of 31 items assessing racial attitudes culled from publicly available databases within each state where each study in the meta-analysis took place. They found that studies involving majority-Black youth had significant lower effect sizes in states with higher levels of anti-Black cultural racism ($g = 0.19$). The authors concluded that "identity-threatening environments themselves may undermine treatment gains" for Black youth (p. 6).

The Present Study

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. The DEMYth-RT theory and Bor et al. (2018) findings led us to predict that the Michael Brown shooting would have an adverse effect on the well-being of all Black youth and teachers in study; in particular, DEMYth-RT would predict that the shooting would increase Black student distractibility and interfere with learning. Likewise, the findings from the Price et al. (in press) study suggest that Black youth would be less likely to benefit from the intervention after the event, especially compared to White youth. However, unlike the Price study which examined the moderated effect of cultural racism attitudes on the response to psychotherapy for Black youth, the present study examined moderated effects of a teacher-delivered intervention on Black youth in the context of a cultural racist traumatic event. Given well-documented evidence that structured, predictable environments and routines may mitigate the harm of traumatic experiences (Kiser et al., 2010), the alternate hypothesis was also viable; that Black students in CHAMPS classrooms may have benefited even more from the intervention after the traumatic event compared to Black youth in the less structured comparison classrooms. Given competing hypotheses were equally likely to be correct based on existing theory and research, we began with a series of six research questions building from examining the main effect to the three-way (event X race X intervention) moderating effects of the event on student and teacher outcomes and differential treatment responses (see Figure 1):

Figure 1. A Simplified Conceptual Framework for Studying the Moderator Effects of Historical Events in Intervention Studies



Note. RQ 1: Path b; RQ 2: Arrow d; RQ 3: Arrow e; RQ 4: Arrow f; RQ 5: Arrow g (student's race as a moderator). RQ 6: Arrow g (teacher's baseline as a moderator).

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 1 assesses the possibility that the event negatively impacted all teachers and students comparably because of the general trauma and turmoil it created for the community.

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 2 examines whether participants in the two conditions were differentially impacted by the event; for instance, CHAMPS may have buffered the effects of the event on teachers and students regardless of race.

RQ 3: Did Black students or teachers have different responses to the shooting and protests compared to White students and teachers? (Race as a moderator of the event.) RQ 3 assesses whether Black students and teachers, regardless of treatment condition, were negatively impacted by the event relative to White participants.

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 4 examines whether Black vs. White participants differentially benefited from the intervention regardless of the event.

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, a three-way interaction.) RQ 5 examines whether Black students' and teachers' treatment response differed before and after the event relative to White participants; for instance, Black and White participants may have had comparable treatment effect sizes before the event, but the effect sizes for Black participants after the event may have eroded due the race-specific trauma of the event.

RQ 6: Did teachers with higher levels of baseline distress have an adverse treatment response after the event compared to those with more normative levels of distress? (Baseline measures as moderators of the event.) RQ 6 assesses whether any negative impact of the traumatic event lowered treatment effects especially for those predisposed to distress regardless of race.

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 85 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 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 recruited (Fall 2012 and 2013, respectively) before the Michael Brown event (Fall 2014) and Cohort 3 was immediately after the Michael Brown event (Fall 2014).

Student participants were 49.6% female and 74.5% African American, 21.1% White, and 4.4% other race (Hispanic/Latinx, Asian). 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 (91%) 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, postintervention. All preintervention assessments occurred in mid-September to mid-October. Postintervention 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.

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 practices, 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. Control condition teachers were offered the CHAMPS intervention immediately after their period of

 DOI:10.1080/2372966X.2021.2024768

participation in the evaluation component of the project ended. Teachers in both conditions received $75 each time they completed surveys to compensate them for their time and effort.

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.

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 postintervention STOIC ratings for the first cohort, we used baseline the 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 (NICHD ECCRN, 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 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), 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, 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 on a 6-point Likert scale. 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., 2019). 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., 2019). 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., 2018; Herman et al., 2020).

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

 DOI:10.1080/2372966X.2021.2024768

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 Assessment, Inc., 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 Assessment, Inc., 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. Given the relatively small sample size, we coded teacher race into two categories: Black and non-Black.

Analytic Approach

For the data analysis, we used multiple imputation for handling missing data (Schafer & Olsen, 1998). The details of multiple imputation are reported in the supplemental material. 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.

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:

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

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

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Treatment_j + \gamma_{12} Event_j + \gamma_{13} Treatment_j{}^{*} Event_j.$$

$$\beta_{qj} = \sum_{q=2}^{Q} \gamma_q$$

Y_{ij} is the student outcome variable for a student i in class j. $Black_{ij}$ is a binary variable that represents whether a student i in class j is Black 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, other race, 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 Michael Brown event ($Event$ = 0 for before the Michael Brown event, and = 1 for after the Michael Brown event). σ^2 and τ^2 are variance components for level 1 and level 2 residuals conditional on these variables. The parameter, γ_{01}, indicates the treatment effect for White students before the Michael Brown event. The parameter, γ_{02}, indicates the effect of the Michael Brown event for the White students in the control group. The parameter, γ_{03}, indicates the moderated effect of the Michael Brown event on the intervention for the White students. The parameter, γ_{10}, indicates the Black–White gap for students in the control group before the Michael Brown event (or the Black–non-Black gap for teachers in the control group). The parameter, γ_{11}, indicates 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}, indicates the additional effect of the Michael Brown event for the Black students (or the additional B–W gap for White students after the Michael Brown event). The parameter, γ_{13}, indicates the additional treatment effect for the Black students after the Michael Brown event (or the additional B–W gap for Black students after the Michael Brown event). The effect sizes (d) for main and moderation effects are calculated by standardizing the corresponding parameters by the pooled standard deviation of the outcome in the unconditional model without including any covariates. For example, the effect size of the treatment effect for White students before the Michael Brown event is calculated by $d = \gamma_{01} / \sqrt{\sigma_0^2 + \tau_0^2}$, where σ_0^2 and τ_0^2 are the variance components for level 1 and level 2 residuals in the unconditional model. The traditional, commonly used small-medium-large distinctions for interpreting effect sizes (e.g., Cohen's 1988 guidance) are not very useful for decision-makers and can be misleading, Bloom et al. (2008), Dong et al. (2016), 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. Note that we tested the full model and we also tested simplified models by removing nonsignificant interaction terms. We tested a total 19 models for five outcome variables and reported the results of the five final models below.

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 standardized mean differences 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.

Table 1. Covariate Balance Checking Among Four Treatment-by-Event Subgroups for the Analytic Sample of Social Behavioral Outcomes at Baseline

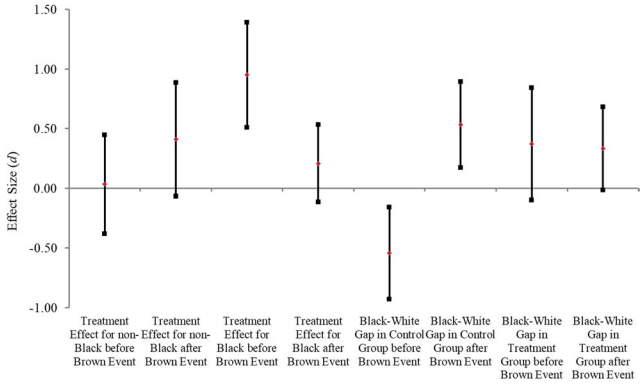
Treatment Status	Control		Control		Treatment		Treatment		Maximum Standardized Mean Difference
Brown Event	Before		After		Before		After		
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Age (year)	12.80	0.88	12.31	0.87	12.47	0.84	12.35	0.86	0.56
Female	0.46	0.50	0.48	0.50	0.54	0.50	0.46	0.50	0.16
Lunch status	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	Point Estimate	SE	*p*-Value
Fixed Effect (γ)			
Intercept	2.23	0.28	<0.0001
Time	−0.04	0.03	0.2812
CLASS-S climate	0.25	0.06	<0.0001
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	<0.0001
Treatment * Event * Black	−0.66	0.23	0.0048
Random Effect (variance)			
Teacher	0.049	0.022	0.0111
Time	0.239	0.026	<0.0001

Note. Akaike's Information Criteria (AIC) = 424.1. Bayesian Information Criteria (BIC) = 456.0. The conditional intraclass correlation coefficient (ICC) is 0.171 and the unconditional ICC is 0.347.

Figure 2. Point Estimates and 95% Confidence Intervals of Treatment Effects and Black–White Gaps Before and After the Event on Teacher STOIC Scores



Note. Effect size is calculated by standardizing the treatment effect or gap by the pooled standard deviation of the outcome.

Event Effects on Teacher Adherence to CHAMPS

The fixed effects of a two-level HLM of STOIC ratings at the end of the school year, adjusting for baseline climate scores are reported in Table 2. It revealed that there was no significant intervention effect for non-Black teachers before the event ($\gamma = 0.02$, $p = 0.87$), no significant event effect for the non-Black teachers ($\gamma = -0.14$, $p = 0.30$; RQ 1), and no moderated treatment effect for the event ($\gamma = 0.22$, $p = 0.25$; RQ 2). However, there was a significant difference on event effect between Black teachers and non-Black teachers ($\gamma = 0.64$, $p < 0.0001$; RQ 3), significant moderated treatment effects for Black teachers before the event ($\gamma = 0.54$, $p = 0.0031$; RQ 4), and significant moderated treatment effects for Black teachers after the event ($\gamma = -0.66$, $p = 0.0048$; RQ 5) (Table 2). We further present the point estimates and 95% confidence intervals regarding treatment effect sizes on different groups and the Black–non-Black gap (Black effect size minus non-Black effect size) under different conditions in Figure 2 and Table S1. 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–non-Black gap in the control group before the event was significant ($d = -0.54$, $p = 0.0061$), the Black–non-Black gap in the control group after the event was also significant ($d = 0.53$, $p = 0.0041$), while the Black–non-Black gaps in the treatment group 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–non-Black gap in classroom management skills favored non-Black teachers before the event, but favored Black teachers after the event. The intervention ultimately made the Black–non-Black gap nonsignificant.

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 S2) 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 ($\gamma = 7.75, p = 0.2209$), no significant event effect for the non-Black teachers (RQ 1), no moderated treatment effect for the event ($\gamma = 11.28, p = 0.3073$; RQ 2), and no significant moderated treatment effect for Black teachers before the event ($\gamma = 13.01, p = 0.2823$; RQ 4). However, there was a significant event effect for Black teachers ($\gamma = 39.13, p = 0.0001$; RQ 3), which suggested that Black teachers in the control group after the event were 39.13 points higher than Black teachers in the control group before the event on the BCIO-R positive-to-negative ratios. There was a significant moderated treatment effect for Black teachers after the event ($\gamma = -68.08, p = 0.0002$; RQ 5), which suggested that Black teachers in the treatment group after the event had 68.08 points lower than Black teachers in the control group after the event on the BCIO-R positive-to-negative ratios. We further present the point estimates and 95% confidence intervals regarding treatment effect sizes on different groups and Black–non-Black gap under different conditions in Table S3 and Figure S1. 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 teachers before the event is not significant ($d = 0.26, p = 0.2209$); the Black–non-Black gap in the control group after the event is significant ($d = 1.00, p < 0.0001$), the Black–non-Black gap in the treatment group after the event is also significant ($d = -0.83, p = 0.0002$), while the Black–non-Black 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–non-Black 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

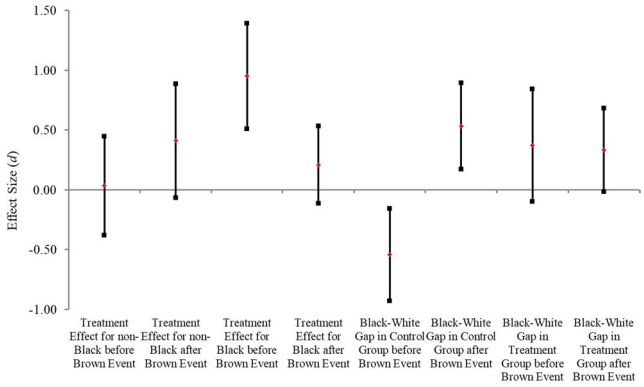
The baseline distress measures did not moderate event-related treatment effects on any teacher or student outcomes (RQ 6). The Michael Brown event or being Black did not have any significant moderation effect on the teacher-reported concentration problems, teacher-reported disruptive

Table 3. HLM Results for 2-Level Model Examining the Effects of CHAMPS on Student Prosocial

Variable	Point Estimate	SE	*p*-Value
Fixed Effect (γ)			
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
Random Effect (variance)			
Teacher	0.046	0.014	0.0006
Student	0.440	0.020	<0.0001

Note. TOCA = Teacher Observation of Classroom Adaptation-Checklist. DBR = Direct Behavior Rating. Akaike's Information Criteria (AIC) = 2255.5. Bayesian Information Criteria (BIC) = 2260.4. The conditional intraclass correlation coefficient (ICC) is 0.094 and the unconditional ICC is 0.309.

Figure 3. Point Estimates and 95% Confidence Intervals of Treatment Effects and Black–White Gaps Before and After the Event on Student TOCA-C Prosocial Behaviors



Note. Effect size is calculated by standardizing the treatment effect or gap by the pooled standard deviation of the outcome.

behavior problems, or teacher-reported emotional dysregulation (RQs 2 & 4). There were some moderation effects on prosocial behavior (Table 3). For instance, the differential impact of the event on prosocial behavior between Black and White students was significant ($\gamma = -0.32$, $p = .014$; RQ 4). Additionally, the treatment effect on Black students after the event was significant ($\gamma = 0.37$, $p = .019$; RQ 5). We further present the point estimates and 95% confidence intervals regarding treatment effect sizes on different groups and the Black–White gap under different conditions in Table S4 and Figure 3. For example, the treatment effect for Black students after the event is significant ($d = 0.40$, $p = 0.007$) while the treatment effect for White students before the event is not significant ($d = 0.04$, $p = 0.51$); the Black–White gap in the control group after the event is significant ($d = -0.34$, $p = 0.002$) while the Black–White gaps under other conditions are not significant. These results indicate that CHAMPS had a moderate and 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 White students.

Differential Effects on Student Academic Outcomes

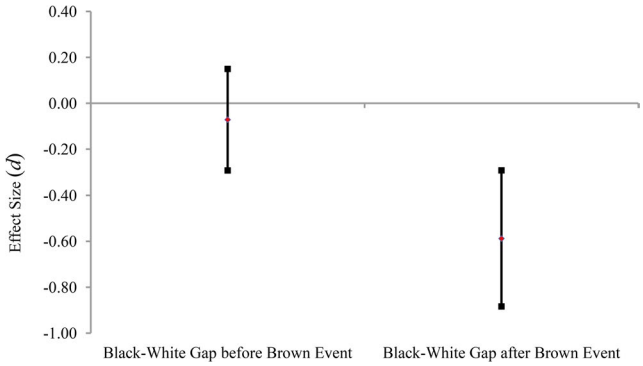
There were no significant moderation effects on the MAP Math or SAT-10 Reading Comprehension scores. However, the event had a significant additional effect on Black students ($\gamma = -17.58$, $p = 0.0153$; RQ 3) on SAT-10 Problem Solving scales (Tables S5). Table S6 and Figure 4 illustrates the Black–White 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 White students before and after the event and Black students before the event on MAP Communication Arts ($d = 0.28$, $p = 0.0002$; Table S8) and intervention had a significant additional negative effect on Black students after the event ($\gamma = -0.39$, $p = 0.0005$; RQ 5; Table S7), which results in a non-significant treatment effect for Black students after the event ($d = -0.19$, $p = 0.1562$) (Table S8; Figure S2).

DISCUSSION

The findings suggest that a traumatic historical event, the shooting of Michael Brown, may have 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 benefited 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 opportunity gap favoring White students significantly increased after the event. In particular, although the intervention had a significant and small benefit for both White 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 findings are consistent with the DEMYth-RT theory which hypothesizes that ongoing experiences with racism undermine Black youth well-being regardless of contextual supports. In particular, during the middle school years, DEMYth-RT emphasizes how these

Figure 4. Point Estimates and 95% Confidence Intervals of Black–White Gaps Before and After the Event on Student SAT-10 Problem Solving



Note. Effect size is calculated by standardizing the gap by the pooled standard deviation of the outcome.

experiences can disrupt identity development and sense of personal safety which can interfere with learning. In the present study, we found evidence that a proximal cultural racist traumatic event, the shooting of Michael Brown, may have undermined Black student academic benefit from a structured teacher-delivered intervention. Prior to the study, we had speculated that the structured, predictable, and nurturing environments that were part of the CHAMPS intervention may minimize or mitigate the damaging effects of the traumatic event on Black youth academic performance. Instead, it appears at least some cultural racist events and experience may be so stressful and threatening to identity that positive and structured environments alone cannot overcome them.

On the other hand, in addition to the apparent negative effect of the event on Black student achievement, findings also revealed a potential intervention benefit for Black students. CHAMPS had a significant and moderate 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.

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–non-Black 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.

LIMITATIONS

The present study design is best described as a natural experiment; that is intervention participants were randomly exposed or not to the historical event during their

participation based on the proximity of their cohort to the event. Although a strong quasi-experimental design, it does not constitute a true experiment and cannot fully rule out alternate explanations for study effects. Our covariate balance check did not reveal large differences on the measured baseline covariates among four treatment-by-moderator subgroups (see Table 1), and we controlled all the covariates in our analysis; however, there may be potential hidden bias due to omitted variables confounding with the event. For instance, it is possible training or coaching differed after the event compared to before it. This possibility is reduced by the fact that the trainer and the coach were the same individuals at each time point, and we did not observe variability in training or coaching implementation quality over time. It is also possible that observers and raters might have been affected by the historical event and provided differential ratings as a result. However, the observers were largely the same individuals at each time point and fidelity ratings were nearly identical before and after the event; that is, there was a high level of agreement among a diverse group of observers at each timepoint and no discernible, systematic bias. Similarly, the historical event might have also affected teachers' student ratings; however, the consistency of effects and the observed differences on student performance measures suggest any teacher-specific rating effects were modest at best.

Another consideration is we did not explore alternate research designs and questions; for instance, if we were principally interested in whether effects from one cohort generalized to other cohorts or whether effects were replicated over time we would have drawn upon theoretical work of Steiner et al. (2019) and others. These alternate designs would answer different research questions—such as how much variability is observed in intervention effects within underpowered RCTs (within each cohort)—than we hypothesized and also would not allow for the examination of effects within treatment conditions as we examined in this study. A strength of our design is that we aggregated across two cohorts prior to the historical effect to increase power and minimize variation observed between cohorts (e.g., cohort one and two were recruited from different schools in the same district but with very different student demographics).

IMPLICATIONS

The findings speak to the power of contexts to influence educational outcomes for students. A prior study revealed the benefit of the program on average for all teachers and students (Herman et al., in press). 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 opportunity gaps in these contexts. Although the intervention improved outcomes for both Black and White students prior to the shooting, academic performance disparities between White and Black students increased over the course of a full academic year in the aftermath of the Michael Brown shooting. 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 (Manzoni, in press; Price et al., in press; Yeager & Walton, 2011). 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 intervention.

The findings have 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). The finding that CHAMPS had a particularly strong benefit for Black teachers prior to the Michael Brown shooting ($d = 0.95$) suggest that this intervention and others like it that focus 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. This adds to the existing literature which suggests that daily experiences of high levels of cultural racism can interfere with social behavioral intervention effectiveness for Black youth (Price et al., 2021; Price et al., in press). 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). Indeed, police brutality incidents against Black citizens are rampant in the U.S. Our findings suggest youth in communities near these events may be benefitting less from social behavioral and educational interventions than their

peers in the aftermath of these events. It is important for investigators to carefully examine and document these contextual features in their reports (Kaplan et al., 2020).

DISCLOSURE

The authors have no conflicts of interest to report.

FUNDING

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130143 to the University of Missouri (PI: Keith Herman). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID

Keith C. Herman https://orcid.org/0000-0003-2246-5792

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.

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. https://doi.org/10.1080/19345740802400072

Bor, J., Venkataramani, A. S., Williams, D. R., & Tsai, A. C. (2018). Police killings and their spillover effects on the mental health of black Americans: A population-based, quasi-experimental study. *The Lancet*, *392*(10144), 302–310. https://doi.org/10.1016/S0140-6736(18)31130-9

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), 297–312. https://doi.org/10.1037/h0040950

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*(4), 195–200. https://doi.org/10.1177/1534508409340391

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 planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, *40*(4), 334–377. https://doi.org/10.1177/0193841X16671283

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.

Haeny, A. M., Holmes, S. C., & Williams, M. T. (2021). The need for shared nomenclature on racism and related terminology in psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(5), 886–892.

Harcourt Assessment, Inc. (2004). *Stanford achievement test series—Tenth edition technical data report*.

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. https://doi.org/10.1177/1098300717732066

Herman, K.C., Prewett, S., Eddy, C., Savale, A., & Reinke, W. M. (2020). Patterns of middle school teacher stress and coping: Concurrent and prospective correlates. *Journal of School Psychology*, *78*, 54–68.

Herman, K. C., Reinke, W. M., Dong, N., & Bradshaw, C. P. (in press). Can effective classroom behavior management increase student achievement in middle school? Findings from a group randomized trial. *Journal of Educational Psychology*.

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. https://doi.org/10.1111/j.1750-8606.2008.00061.x

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. *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *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. https://doi.org/10.3102/0013189X20921862

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), 86–95. https://doi.org/10.1037/spq0000263

Kiser, L. J., Medoff, D. R., & Black, M. M. (2010). The role of family processes in childhood traumatic stress reactions for youths living in urban poverty. *Traumatology*, *16*(2), 33–42. https://doi.org/10.1177/1534765609358466

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.

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. https://doi.org/10.1177/0748175609333560

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*(9), 606–613.

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). National Center for Special Education Research.

Manzoni, A. (in press). Equalizing or stratifying? Intergenerational persistence across college degrees. *The Journal of Higher Education*.

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

McLean, C. P., Handa, S., Dickstein, B. D., Benson, T. A., Baker, M. T., Isler, W. C., Peterson, A. L., & Litz, B. T. (2013). Posttraumatic growth and posttraumatic stress among military medical personnel. *Psychological Trauma: Theory, Research, Practice, and Policy*, *5*(1), 62.

Missouri Department of Elementary and Secondary Education. (2015). *Missouri assessment of progress test*. 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., 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), 13–27. https://doi.org/10.1037/a0018576

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

Price, M. A., McKetta, S., Weisz, J. R., Ford, J. V., Lattanner, M. R., Skov, H., Wolock, E., & Hatzenbuehler, M. L. (2021). Cultural sexism moderates efficacy of psychological therapy for girls: Results from a spatial meta-analysis. *Clinical Psychology: Science and Practice*, *28*(3), 299.

Price, M. A., Weisz, J. R., McKetta, S., Hollinsaid, N. L., Lattanner, M. R., Reid, A. E., & Hatzenbuehler, M. L. (in press). Meta-analysis: Are psychotherapies less effective for Black youth in communities with higher levels of anti-Black racism? *Journal of the American Academy of Child & Adolescent Psychiatry*.

Priest, N., Paradies, Y., Trenerry, B., Truong, M., Karlsen, S., & Kelly, Y. (2013). A systematic review of studies examining the relationship between reported racism and health and wellbeing for children and young people. *Social Science & Medicine*, *95*, 115–127. https://doi.org/10.1016/j.socscimed.2012.11.031

Reid, A. E., Dovidio, J. F., Ballester, E., & Johnson, B. T. (2014). HIV prevention interventions to reduce sexual risk for African Americans: The influence of community-level stigma and psychological processes. *Social Science & Medicine (1982)*, *103*, 118–125. https://doi.org/10.1016/j.socs-cimed.2013.06.028.

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. https://doi.org/10.1007/s11121-018-0932-3

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. https://doi.org/10.1177/1098300715570640

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. https://doi.org/10.1177/0042085912454441

Saleem, F. T., Anderson, R. E., & Williams, M. (2020). Addressing the "myth" of racial trauma: Developmental and ecological considerations for youth of color. *Clinical child and family psychology review*, *23*(1), 1–14.

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.

Sharkey, P., Schwartz, A. E., Ellen, I. G., & Lacoe, J. (2014). High stakes in the classroom, high stakes on the street: The effects of community violence on student's standardized test performance. *Sociological Science*, *1*, 199–220. https://doi.org/10.15195/v1.a14

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

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

Steiner, P. M., Wong, V. C., & Anglin, K. L. (2019). A causal replication framework for designing and assessing replication efforts.. *ZeitschriftFürPsychologie / Journal of Psychology*, *227,* 280–292. https://doi.org/10.1027/2151-2604/a000385

Strickler, R., & Lawson, E. (2020). Racial conservatism, self-monitoring, and perceptions of police violence. *Politics, Groups, and Identities*, 1–22. https://doi.org/10.1080/21565503.2020.1782234

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. In M. J. Friedman, T. M. Keane, & P. A. Resick (Eds.), *Handbook of PTSD: Science and practice* (pp. 99–115). Guilford.

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

AUTHOR BIOGRAPHICAL STATEMENTS

Keith C. Herman is a Curator's Distinguished Professor in Department of Education, School, & Counseling Psychology at the University of Missouri. He is the Co-Founder and Co-Director of the Missouri Prevention Science Institute. He has an extensive grant and publication record including over 150 peer-reviewed publications in the areas of prevention and early intervention of child emotional and behavior disturbances and culturally-sensitive education interventions.

Nianbo Dong is an Associate Professor and Chair in the Area, Learning, Development, and Psychological Sciences at the University of North Carolina, Chapel Hill. Dr. Dong's research program centers on developing and applying rigorous quantitative methods to evaluate educational policies, programs, and practice. His current interests in quantitative methodology focus on power analyses of the main, moderation, and mediation effects in multilevel experiments and causal inference. His substantive

research focuses on the evaluations of the effectiveness of teacher and principal training programs and early child education programs. His work has been supported by over $5 million of funding as the Principal Investigator or Co-Principal Investigator. He also received the NSF Faculty Early Career award in 2017.

Wendy M. Reinke is a Professor in the Department of Education, School, & Counseling Psychology at the University of Missouri. She is the Co-Founder and Co-Director of the Missouri Prevention Science Institute. She has an extensive grant and publication record including over 100 peer-reviewed publications and over $50 million in grant funding in the areas of prevention and early intervention of child emotional and behavior disturbances. She is also the Director of the National Center for Rural School Mental Health and the co-developer and leadership team member for the Family Access Center of Excellence and the Boone County Schools Mental Health Coalition.

Catherine P. Bradshaw is a Professor and the Associate Dean for Research and Faculty Development at the School of Education and Human Development at the University of Virginia. Prior to her current appointment at U.Va., she was an Associate Professor and the Associate Chair of the Department of Mental Health at the Johns Hopkins Bloomberg School of Public Health, where she maintains an adjunct faculty position. She holds a doctorate in developmental psychology from Cornell University and a master's of education in counseling and guidance from the University of Georgia.Her primary research interests focus on the development of aggressive behavior and school-based prevention. She collaborates on research projects examining bullying and school climate; the development of aggressive and problem behaviors; effects of exposure to violence, peer victimization, and environmental stress on children; children with emotional and behavioral disorders and autism; and the design, evaluation, and implementation of evidence-based prevention programs in schools. She has led a number of federally funded randomized trials of school-based prevention programs, including Positive Behavioral Interventions and Supports (PBIS) and social–emotional learning curricula. She also has expertise in implementation science and coaching models.