Administration and Policy in Mental Health and Mental Health Services Research (2020) 47:972–986
https://doi.org/10.1007/s10488-020-01034-1

ORIGINAL ARTICLE

Using Complier Average Causal Effect Estimation to Examine Student Outcomes of the PAX Good Behavior Game When Integrated with the PATHS Curriculum

Catherine P. Bradshaw[1] · Kathan D. Shukla[2] · Elise T. Pas[3] · Juliette K. Berg[4] · Nicholas S. Ialongo[3]

Published online: 15 April 2020

Abstract
A growing body of research has documented a link between variation in implementation dosage and outcomes associated with preventive interventions. Complier Average Causal Effect (CACE; Jo in J Educ Behav Stat 27:385–409, 2002) analysis allows for estimating program impacts in light of variation in implementation. This study reports intent-to-treat (ITT) and CACE findings from a randomized controlled trial (RCT) testing the impacts of the universal PAX Good Behavior Game (PAX GBG) integrated with Promoting Alternative Thinking Strategies (i.e., PATHS to PAX) and PAX GBG only compared to a control. This study used ratings by 318 K-5 teachers of 1526 at-risk children who, at baseline, were rated as displaying the top 33rd percentile of aggressive-disruptive behavior. Leveraging a prior study on these data (Berg et al. in Admin Policy Ment Health Ment Health Serv Res 44:558–571, https://doi.org/10.1007/s10488-016-0738-1, 2017), CACE was defined as the effect of intervention assignment for compliers, using two compliance cut points (50th and 75th percentile), on posttest ratings of student academic engagement, social competence, peer relations, emotion regulation, hyperactivity, and aggressive-disruptive behavior. The ITT analyses indicated improvements for students in the integrated condition on ratings of social competence compared to the control condition. The CACE analyses also indicated significant effects of the integrated intervention on social competence, as well as academic engagement and emotion regulation for students in high compliance classrooms. These findings illustrate the importance of considering variation in implementation within the context of RCTs.

Keywords Prevention · Schools · Implementation · Causal inference · Randomized controlled trial

Introduction

Universal preventive interventions are widely used in schools with the goal of improving a variety of student academic and behavioral outcomes. Despite increased interest and uptake in such preventive interventions, ensuring high fidelity of these interventions, particularly when used in real-world settings, remains a challenge (Domitrovich et al. 2008). Such implementation challenges are also often encountered in school-based research (e.g., see Durlak and Dupree 2008; Durlak et al. 2011; Fixsen et al. 2005). Yet the majority of published studies on school-based trials have employed an intent-to-treat (ITT) approach, whereby the researchers estimated the effect of an intervention based on assignment to intervention condition (Schochet et al. 2014), with an assumption of full implementation for the full sample of participants. Effects for non- or poor-implementers are often small or null, and thus the ITT estimates may understate the effects of the intervention when implemented

✉ Catherine P. Bradshaw
cpb8g@virginia.edu

Kathan D. Shukla
kathans@iima.ac.in

Elise T. Pas
epas@jhu.edu

Juliette K. Berg
jberg@air.org

Nicholas S. Ialongo
nialong1@jhu.edu

1 Curry School of Education and Human Development, University of Virginia, Charlottesville, VA, USA

2 Indian Institute of Management, Ahmedabad, Gujarat, India

3 Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

4 American Institutes for Research, Washington, DC, USA

as intended (Stuart et al. 2008). Taking into consideration the common variation in implementation compliance can be accomplished through an alternative analysis called the complier-average causal effect (CACE; Jo 2002), which estimates treatment effects accounting for variability in implementation fidelity (see Angrist et al. 1996; Little and Yau 1998; Jo 2002; Stuart et al. 2008). More specifically, CACE has been used to estimate the effects of preventive interventions, while accounting for noncompliance, in several randomized studies including children and families (e.g., Barnard et al. 2003; Berg et al. 2017; Connell et al. 2007; Stanger et al. 2011).

The current study builds upon recent findings of a CACE study by Berg et al. (2017), which focused on a classroom-based preventive intervention implemented by teachers called the PAX Good Behavior Game (PAX GBG; see Bradshaw et al. 2009; Ialongo et al. 1999; Ialongo et al. 2001; Kellam et al. 1998, 2008). This earlier study by Berg et al. reported significant impacts of the intervention on teachers' self-efficacy, however, the effects on burnout were less favorable in high compliance classrooms, when the PAX GBG program was integrated with a social-emotional learning curriculum called Promoting Alternative Thinking Strategies (PATHS; Greenberg et al. 2011; Kusché et al. 2011). The current study leveraged the implementation compliance data from that same randomized controlled trial (RCT; see Berg et al. 2017) to estimate the impact of both the integrated (i.e., PATHS to PAX) and PAX GBG only programs, with regard to student outcomes.

Estimating Compliance in School-Based Program Implementation

Teachers are often the primary implementers of classroom-based preventive interventions, yet the degree to which they implement the intervention often varies considerably (Domitrovich et al. 2008), which in turn attenuates program impacts (Durlak and Dupree 2008; Durlak et al. 2011). Implementation compliance is defined as "the discrepancy between what is planned and what is actually delivered when an intervention is conducted" (Domitrovich et al. 2008, p. 7, also see Chen 1998; Hulleman and Cordray 2009; O'Donnell 2008). A common indicator of compliance of school-based interventions is program dosage, which includes the frequency with which or time dedicated to program implementation, with the expectation that a higher dosage would be associated with better outcomes. The dosage of an intervention is relatively straightforward in some manualized curricula or programs, where there are a set number of sessions to implement and an expectation that all are completed, but is more complex with an intervention like the PAX GBG where there are no predefined number of "sessions" to conduct. Such preventive interventions have been shown to vary in implementation dosage (e.g., see Domitrovich et al. 2015) and thus intervention effects can be impacted. Yet relatively few studies have systematically assessed intervention impacts under varying levels of implementation dosage compliance. Additional research is needed to examine the causal impact of teacher-led programs on student outcomes, while taking into consideration program implementation.

As noted earlier, Complier Average Causal Effect (CACE; Jo 2002) analysis is a causal inference analytic approach that estimates treatment effects, accounting for levels of implementation compliance. Although CACE analyses have been conducted in the context of preventive intervention trials (e.g., Barnard et al. 2003; Connell et al. 2007; Stanger et al. 2011), there has been less focus on classroom-based preventive interventions implemented by teachers. In the current study, we used CACE to estimate the effects of the classroom-based, teacher-implemented interventions on student outcomes while accounting for teachers' compliance with intervention implementation, which we operationalized as dosage (Berg et al. 2017). We similarly applied the framework used by Angrist et al. (1996), which outlined a process for a two-arm trial with binary compliance in the potential outcomes framework (also see Frangakis and Rubin 2002; Holland 1986). They defined four compliance types based on individuals' treatment assignment status (1 = treatment, 0 = control) and potential treatment receipt status (1 = received/participated, 0 = not received/not participated). These groups are important because we assumed that the treatment and control groups were likely to have the same proportion of each compliance type because of the group randomization. Therefore, the difference between the treatment and control condition within each compliance type can be interpreted as a causal effect (Frangakis and Rubin 2002).

Research Support for the Current Interventions

As noted above, the current study tested two evidence-based elementary school prevention programs: the PAX version of the Good Behavior Game (PAX GBG; Embry et al. 2003) and Promoting Alternative Thinking Strategies (PATHS; Greenberg et al. 2011; Kusché et al. 2011). Specifically, a three-arm randomized controlled trial (RCT) design was used to compare the PAX GBG only and an integration of the PAX GBG and the PATHS program (Domitrovich et al. 2010) to a control group. PAX GBG provides teachers with an efficient way to reinforce the inhibition of aggressive/disruptive and off-task behavior in a game-like context (Embry et al. 2003). Several large RCTs of GBG have demonstrated positive effects on student peer relations, aggressive/off-task

behavior, substance use, and academic outcomes (e.g., Bradshaw et al. 2009; Ialongo et al. 1999, 2001; Kellam et al. 2008). In a complementary approach, the PATHS curriculum trains teachers to promote the development of emotional awareness and communication, self-regulation, social problem solving, and relationship management skills (e.g., interpersonal skills, conflict management) through didactic lessons that take place weekly across the school year (Greenberg and Kusche 2006). Prior RCTs of PATHS have yielded positive effects on student social-emotional skills, peer relations, prosocial cognitive functioning, social competence, and behavioral adjustment (e.g., Conduct Disorder Problems Research Group 1999; Greenberg and Kusche 2006; Greenberg et al. 1995).

Recent ITT findings from the current RCT on student outcomes provided some evidence that the integrated condition produced slightly more favorable effects for the PATHS to PAX condition relative to PAX GBG only (Ialongo et al. 2019). Specifically, one main intervention effect emerged in these analyses, and that was for the integrated PATHS to PAX condition relative to controls on problem behavior; however, the effect size was small (i.e., Cohen's $d = 0.08$). In addition, students in the PAX GBG only condition with elevated baseline problem behaviors experienced significant improvements relative to controls on problem behaviors. Similarly, students with the lowest baseline teacher–rated readiness to learn and social competence experienced the greatest growth in these outcomes following exposure to the integrated intervention relative to controls. These moderated intervention effects are consistent with prior studies of GBG (Kellam et al. 1998) and a number of other universal prevention programs (e.g., Positive Behavioral Interventions and supports; PBIS, Bradshaw et al. 2015), suggesting that universal program effects are often more pronounced for children who were at-risk at baseline.

Although most prior research on PAX GBG and PATHS individually has focused on student outcomes, there is also growing interest in the impact of these and other such programs on teacher outcomes. For example, using an ITT approach on data from the current RCT, Domitrovich et al. (2016) indicated that teachers in the integrated condition reported feeling more efficacious and feeling more personal accomplishment relative to control teachers after the intervention. As mentioned above, prior work by Berg et al. (2017) using CACE analyses on these same data revealed similar but stronger program effects on personal accomplishment among teachers most likely to comply in both the PAX GBG and integrated intervention conditions. Whereas the ITT analyses demonstrated no significant differences between treatment and control in emotional exhaustion or depersonalization when accounting for implementation compliance, Berg et al. actually found *elevated* emotional exhaustion among higher (i.e., more compliant) implementers of the integrated program. Further, all effects were notably stronger among those meeting the high implementation cut point, with the exception of depersonalization. PATHS to PAX teachers, on average, had greater increases in efficacy as compared to the control condition. The effects for behavior management efficacy seemed to be concentrated among higher implementing teachers, whereas effects did not vary based on implementation level for social-emotional efficacy. Taken together, the findings of the Berg et al. study suggested that important variation in compliance was functionally associated with teacher outcomes; this may also be true for student outcomes, which is the focus of the current study.

Current Study

To date, few studies have tested the impacts of either PAX GBG or PATHS on students when variation in teacher implementation is taken into consideration. Specifically, PAX GBG is largely a behavior management program focused on teaching and reinforcing inhibition, whereas PATHS is a social-emotional learning program aimed at fostering a broader set of skills related to self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. Thus, the integration of the two models may be particularly synergistic and address a broader range of skills children need to be successful at school and in life (Domitrovich et al. 2008). Prior research would suggest that higher compliance in implementing either model would likely result in more positive effects for students (Durlak and Dupree 2008). Toward that end, this study compared parallel ITT analyses to CACE analyses as a means for estimating the impacts of the integrated version of the PAX Good Behavior Game and Promoting Alternative Thinking Strategies (i.e., PATHS to PAX), relative to PAX GBG only and a control, on student outcomes, as rated by teachers over one school year.

We leveraged prior CACE research by Berg et al. examining teacher intervention compliance at a medium and high dosage level in relation to teacher outcomes to estimate impacts on teacher ratings of student academic engagement, social competence, peer relations, emotion regulation, hyperactivity, and aggressive-disruptive behavior, controlling for student demographics and baseline teacher ratings. In light of prior research on the GBG (Kellam et al. 1998, 2014) suggesting that baseline risk moderated intervention effects, we examined the outcomes for the higher-risk students. Thus, we selected a subset of students rated by teachers at baseline as demonstrating elevated aggressive disruptive behavior scores (i.e., top 33rd percentile) as the subsample of interest with whom to test parallel ITT and CACE effects for both interventions relative to control.

Using data on this subsample of students, we examined CACE and focused on behavioral outcomes at the student level in the context of a nested design in which implementation compliance occurred at the classroom level. Further, we systematically explored classroom-level implementation variation (i.e., medium and high compliance) and directly compared the findings to those of a parallel ITT analysis, which did not account for compliance. A novel aspect of the study design was the planned contrast of the PAX GBG classroom management model when implemented alone to that of an integrated intervention, combining PAX GBG with the PATHS social-emotional learning program. We operationalized implementation compliance as the teachers' use of the PAX GBG "games" in the classroom, using records of how many games they played throughout the school year and for how long they played each game. More specifically, compliance was defined as being above a cut point on both the number of games played and the total number of minutes of games played. Consistent with prior research suggesting a link between implementation dosage and stronger outcomes (e.g., Durlak and Dupree 2008), we expected to find stronger effects for students in classrooms of intervention teachers who sufficiently complied with implementing at a high dosage. Based on our prior ITT findings (Domitrovich et al. 2016) and CACE analyses (Berg et al. 2017) regarding teacher and student outcomes (Ialongo et al. 2019), we also anticipated that student effects would be most pronounced in the integrated PATHS to PAX condition relative to both the control and PAX GBG only conditions.

Method

Sample

Teachers

As in the original study by Berg et al. (2017), we used data from the 27 elementary school RCT, which included 350 K-5 teachers. Schools, and therefore teachers, were enrolled in three cohorts (i.e., for one year each, in three consecutive years) and teachers provided consent for their voluntary participation. The sample was generally evenly split across the three cohorts (31% cohort 1, 34% cohort 2, and 35% cohort 3) and across the three conditions (25% PAX GBG, 29% PATHS to PAX, 37% control). The vast majority of the teacher sample was female (i.e., 88%). Less than half of the teachers were 30 years old or younger (41.4%) and taught grades 3 through 5 (44.1%). Just over half of the teachers had attained a graduate degree (56.4%). Due to missing data on the compliance measure, 32 cases (19 in the integrated condition and 13 cases in the PAX GBG condition) were removed from all analyses through listwise deletion, resulting in a total analysis sample size of 318 teachers. Importantly, there was no evidence that missingness was systematically related to other variables of interest or study condition (see Ialongo et al. 2019). See Table 1 for additional details on the sample as well as average scores on the key measures administered in this study.

Table 1. Descriptive statistics for student and teacher participants

	Full sample	Control condition	Integrated PATHS to PAX condition	PAX GBG only condition
Student characteristics in the at-risk subsample ($n = 1526$)				
Free/reduce-priced priced meals (%)	92.50	96.80	90.90	88.80
Gender (%)				
Male	60.10	60.30	57.00	62.30
Female	39.90	39.70	43.00	37.70
Race/ethnicity (%)				
African American	93.70	96.20	93.90	90.60
Hispanic	3.10	2.00	2.80	4.70
White	2.90	1.80	2.80	4.30
Asian	0.20	0.00	0.30	0.40
Multiple	0.10	0.00	0.30	0.00
Special education services (%)	13.40	15.20	16.00	9.70
Teacher characteristics ($n = 350$)				
Female (%)	88.7	88.3	90.2	88.0
Taught grades 3–5 (%)	41.5	42.2	45.1	38.0
Age ≤ 30 years (%)	38.7	32.8	39.0	45.4
Has graduate degree (%)	52.8	50.0	52.4	56.5

Students

Eligible participants included students enrolled in K-5 classrooms in each of the 27 participating schools at the beginning of the school year. Across the three cohorts, there were a total of 7024 students enrolled in the participating schools during our baseline, or pre-test, period. Of the total eligible, we obtained written parent consent for 79.9% ($N = 5611$); 7.1% refused participation and 12.9% did not respond to the consent request. Of the 5611 enrolled students, 50.4% were male, 89.6% were African American, and 86.5% received free and reduce-priced meals. The mean grade level was 2.36. The demographic profiles of the study students were comparable to the overall school profiles in terms of gender, ethnicity and free and reduce-priced meals (FARMs; a proxy for family income); students in the schools were 50.81% male, 87.93% African American, and 85.96% FARMs eligible.

A subsample of the full RCT student sample was selected for inclusion in the current study based on baseline aggressive-disruptive behavior ratings by their teachers. Specifically, we identified the subsample of high-risk students, defined as those having scores within the top 33rd percentile on the aggressive-disruptive behavior scale of the TOCA-R teacher rating (described below). As noted above, we selected this subsample based on prior research documenting that the effects of universal programs, including the GBG, are often most salient among students with the highest level of aggressive behavior (Kellam et al. 1998, also see Bradshaw et al. 2015). This resulted in a subsample of 1526 high-risk students (60.6% male) who were included in the current study and were evenly distributed across the three conditions (34% PAX GBG only, 27% PATHS to PAX integrated, 39% control). The majority of students included in the analyses were African American (93.7% on average) and received free and reduce-priced meals (i.e., FARMs; 92.5%). In addition, based on teacher report, approximately 22% of the subsample of students were referred during that school year for special education assessment. See Table 1 for additional student demographics by condition.

Procedures

The 27 elementary schools were recruited and principals agreed to participate in a randomized controlled trial of two intervention models and to potentially receive one year of training and coaching for implementation. Teachers were actively consented each year, and then schools were randomized (i.e., cluster randomized trial) to one of three conditions: the PAX GBG only (9 schools), the integration of PAX GBG and PATHS [referred to as PATHS to PAX (P2P); 9 schools], and a control condition (9 schools) where teachers conducted their usual practice. Participating teachers and schools received a modest incentive for completion of each wave of data collection; intervention teachers received stipends for attendance at the training. Given the focus on a whole-school approach, principals also encouraged participation. Parents provided consent for the teachers to complete ratings of their child; students who returned a signed consent form, regardless of whether the parents agreed or did not agree to let this child participate, were eligible to participate in a class-wide pizza party. The Institutional Review Board provided approval of this study. For additional details on the training procedures and interventions, see Ialongo et al. (2019).

Measures

All student outcomes were assessed using a teacher-report measure administered twice, first in the fall as a baseline, and in the spring as an end of school year posttest.

Student Demographic Covariates

Baseline student gender, ethnicity (i.e., Black and Hispanic racial/ethnic groups, dummy coded), and student special education status were included in the regression models as covariates.

Student Outcomes

Teachers completed a checklist version of *Teacher Observation of Classroom Adaptation-Revised* (TOCA-R; Bradshaw and Kush 2019; Koth et al. 2009; Werthamer-Larsson et al. 1991). The TOCA-R required teachers to rate the child's adaptation to classroom task demands over the last three weeks across six scales on a 6-point frequency scale (1 = *never* to 6 = *almost always*). The domains were: academic engagement (3 items; i.e., completed assignments, learned up to ability, and eager to learn; Cronbach's alpha [$\alpha = .89$]); social competence (8 items, e.g., resolves peer problems on his/her own, expresses feelings appropriately, and showed empathy and compassion for others' feelings; $\alpha = .94$), positive peer relations (3 items, i.e., liked by classmates, other children sought him/her out to play, and disliked by classmates; $\alpha = .83$), emotion regulation (4 items, e.g., controlled temper when there was a disagreement, could calm down when excited or all wound up, and coped well with disappointment or frustration; $\alpha = .88$), inattention/hyperactivity (6 items, e.g., paid attention, stays on task, and concentrated on class work; $\alpha = .87$); and aggressive-disruptive behavior (15 items, e.g., lied, started physical fights, stubborn, broke rules, hurt others physically, and yelled at others; $\alpha = .96$). All items were scored such that a higher score indicated more of that construct, which required some individual items to be reverse scored. All outcome measures

were completed by teacher-report at two time points: fall baseline and spring posttest, at the end of the school year. Descriptive statistics of student demographics for each condition are reported in Table 1. Table 2 presents descriptive information on baseline and posttest scores of outcomes by intervention condition. Several studies have documented the reliability, validity and psychometric properties of the TOCA-R and its various subscales (e.g., Bradshaw and Kush 2019; Bradshaw et al. 2015; Ialongo et al. 2019; Koth et al. 2009; Werthamer-Larsson et al. 1991).

Compliance

This paper focuses largely on examining compliance in relation to the student outcomes; therefore, we refer readers to the original study by Berg et al. (2017) for additional details on the CACE modeling of teacher compliance. In both studies, compliance was assessed using intervention teachers' weekly data logging the number of games they played and the number of minutes they spent on each game. The sum of each of the number of games and minutes played across the school year served as the total games and minutes played. There were no significant differences between the two conditions with respect to the number of games and minutes played, F (14), 0.08, $p > 77$ for games played; F (14), 0.46, $p > .50$ for minutes played; [for additional details, see Ialongo et al. (2019)].

The compliance needed to be dichotomous and thus a cut point needed to be set, but there was a trade-off when choosing this value. For example, if the cut point was 200 games, the assumption would be that students in classrooms where the teacher led less than 200 games would not be affected by the intervention. However, setting the cut point too low would lead to greater variation in the degree to which compliers implemented the program and would not capture a meaningful contrast in compliance levels. Moreover, with a higher cut point, the sample size among these complier outliers becomes small and implies a larger estimated CACE, in turn reducing the utility and interpretability of the CACE estimates. In addition, a higher cut point would consider a large proportion of teachers to be non-compliers, even though they did actually implement the program fairly regularly, albeit to a lesser degree than those at this very high level. As a result, when defining compliance, there was a trade-off when deciding where to set the cut point, so we operationalized compliance two ways: (1) using a *medium compliance* cut point for teachers who fell above the 50th percentile on both the number of games played and the minutes played (i.e., 138 games and 1145 min; $n = 81$ total treatment teachers) and (2) a *high compliance* cut point for teachers who fell above the 75th percentile on both the number of games and minutes played (i.e., 214 games and 2110 min; $n = 29$ total treatment teachers). This is consistent with Berg et al. (2017), where additional details of the

Table 2 Descriptive statistics for teacher-rated student outcomes by intervention condition ($n = 1526$)

	Full Sub-sample		Control condition		Integrated PATHS to PAX condition		PAX GBG only condition	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Academic engagement								
Baseline	3.61	1.13	3.66	1.13	3.60	1.12	3.55	1.13
Posttest	3.67	1.23	3.62	1.18	3.82	1.29	3.63	1.23
Social competence								
Baseline	3.06	0.81	3.10	0.84	3.07	0.81	3.00	0.77
Posttest	3.23	0.97	3.16	0.91	3.41	1.05	3.18	0.97
Peer relations								
Baseline	3.85	0.96	3.80	0.94	3.95	0.99	3.83	0.97
Posttest	3.87	1.02	3.80	0.94	4.07	1.12	3.78	0.98
Hyperactivity								
Baseline	3.54	0.84	3.50	0.85	3.56	0.83	3.58	0.83
Posttest	3.37	0.91	3.36	0.87	3.3	0.97	3.42	0.89
Emotion regulation								
Baseline	3.01	0.89	3.00	0.91	3.07	0.90	2.97	0.86
Post test	3.15	1.07	3.07	0.98	3.31	1.16	3.12	1.08
Aggressive-disruptive								
Baseline	2.97	0.80	3.02	0.82	2.95	0.78	2.93	0.79
Post test	3.02	1.00	3.06	0.98	2.94	1.03	3.03	1.00

Scales range from 1 = *never* to 6 = *almost always*

CACE estimation process, including the assumptions of CACE analyses are provided.

Data Analysis

Assumptions of CACE

Implementation behavior of each participating teacher can only be observed under the condition the teacher is assigned to; thus, CACE cannot be calculated directly comparing outcomes under the treatment and under the control conditions. This is, in turn, the fundamental problem of causal inference (Holland 1986). As a result, it is important to consider a number of assumptions of the CACE analysis, which under certain conditions allow one to identify the causal effect at the average level. In particular, the set of conditions or assumptions outlined by Angrist et al. (1996) have been widely used in CACE analyses and are presented below.

A core assumption is *ignorable treatment assignment,* which provides the basis for causal inference as it guarantees the comparability between treatment arms. In this study, this assumption was automatically satisfied because schools were randomized to conditions. A second assumption is the *stable unit treatment value* (SUTVA), which means that the potential outcome of each individual is not affected by the treatment assignment status of other individuals. This is a questionable assumption within school settings because teachers in the same school are highly likely to interact with one another; however, this concern was minimized in the current trial, as we employed cluster randomization, where the school was the unit of randomization. Previous studies suggested that by employing cluster randomized trials, interaction or contamination among individuals becomes a more manageable problem (Jo et al. 2008a, b; Sobel 2006). Although we cannot prevent interactions among teachers across schools in different intervention arms, the likelihood remains about the same as that observed when no nesting exists. A third assumption is *monotonicity*, which assumes that there are no defiers (i.e., those who do not implement if assigned to the treatment group and do implement if assigned to the control group). This is a reasonable assumption for the control condition, because teachers in the control group did not have access to the intervention. CACE also assumes that there are at least some compliers, meaning that the offer of the intervention induces at least some teachers to implement. This was a reasonable assumption in the current study, particularly given the ongoing coaching provided. Finally, the *exclusion restriction* assumes that always-takers and never-takers in the control group will not have students who benefit from the program and therefore the distribution of outcomes is the same in the treatment and control groups for these two types. Always-takers are those teachers who would always implement the treatment, and never-taker teachers are those who would never implement, regardless of treatment assignment. In our context, this meant that there were no effects of assignment for students in classrooms led by never-takers. Since teachers, and thus their students, in the control group did not have access to the intervention, the stratum of always-takers did not apply to our study.

A nuance of the application of CACE to the current study is that we were estimating the impacts of a school-based intervention delivered by classroom teachers. Therefore, compliance was operationalized at the classroom level, whereas the effects of interest were experienced by students nested within classrooms (i.e., a multilevel framework). Further, the compliers were those teachers who delivered the treatment (i.e., implemented) when assigned to the treatment group and did not deliver the intervention when assigned to the control condition. In an ITT analysis, the effect of treatment assignment is the same as the effect of full participation for the compliers. Similar to Angrist et al. (1996), our primary interest in this study was the causal treatment effect for those students in the compliers' classrooms (i.e., complier-average causal effect; CACE). This focus on implementation within a nested design, coupled with the use of a continuous compliance indicator for which we set a threshold of high and medium compliance, rather than traditionally categorizing compliant vs. non-compliant, together make this application of CACE particularly unique.

Given our simplified setting with only compliers and never-takers, we use non-compliers to refer to never-takers. The assumption of the exclusion restriction may need to be relaxed in school-based interventions where compliance, or dosage in this case, is measured across a continuum, as opposed to being determined dichotomously (i.e., compliant vs. non-compliant). The cut point for determining "compliance" has implications as aforementioned, as any cut point above zero games, as is the goal here, would include teachers in the non-compliers group even if they implemented some games. Thus, the assumption that there would be no impact on students in classrooms defined as having never-taker teachers is harder to meet when selecting a non-zero dosage cut point.

Estimation of CACE

As reported by Berg et al. (2017), the CACE models were estimated separately in M*plus* for each of the treatment conditions relative to control (i.e., PATHS to PAX integrated v. control and PAX GBG only vs. control). Linear regression models for each of the outcomes with baseline scores as covariates were conducted. Other control variables included student gender, referral for special education, and dummy-coded variables for Black and Hispanic racial/ethnic group. CACE was defined as the effect of intervention assignment for compliers on the posttest score of each outcome. As in

Berg et al., compliance with the treatment was defined in the current study using two different cut points (50th and 75th percentile of dosage), as a sensitivity analysis for both the cut point and the potential deviation from the exclusion restriction. Specifically, the CACE models were identified in two different ways. First, we assumed the exclusion restriction. In this model, the student-level outcome was regressed on treatment assignment in the complier class but not in the non-complier class and we were assuming that non-compliers were not affected by treatment assignment. However, this assumption might have been violated in our trial. Given these possibilities of deviation from the exclusion restriction, we additionally conducted CACE estimation assuming that the intervention effects were additive (see Jo et al. 2008a, b; Jo et al. 2008a, b). In the model with the additive treatment effect assumption (i.e., instead of the exclusion restriction), the outcome was regressed on intervention in both the complier and non-complier classes.

As noted earlier, missing data on the compliance measure caused 32 cases (19 in the integrated condition and 13 cases in the PAX GBG condition) to drop out through listwise deletion, resulting in a total sample size of 318 teachers across three arms of the study. An additional set of 34 teachers was excluded in the analysis phase since we restricted the student sample to the top 33 percentile on the aggressive-disruptive behavior and due to missing data on the student covariates. Specifically, in the models comparing the integrated condition to the control condition (i.e., excluding the PAX GBG teachers), 25 teachers were dropped, resulting in a sample size of 185 teachers (i.e., out of an expected 229 teachers without missing data). In the models comparing the PAX GBG condition to the control condition, 34 teachers were dropped, resulting in a total sample size of 202 (i.e., out of the full sample of 249 teachers without missing data). In principle, we could incorporate all cases including the ones with incomplete information. However, we employed listwise deletion, given that there is little research providing guidelines for handling the simultaneous complications of noncompliance, clustering, and missing data. In this study, we focused on the handling of noncompliance and clustering, and ignored biases introduced by dropping teachers with missing data, which is a limitation of the study. We used maximum likelihood (ML) estimation with the expectation maximization (EM) algorithm (Little and Rubin 2002) for CACE estimation, which incorporated the mixture modeling feature in M*plus*. In this framework, compliance status was defined by a dichotomous latent variable, with one class referring to the compliers and the other class referring to the non-compliers. Given our simplified setting where there are only compliers and never-takers or non-compliers, the compliance class membership was completely observed in the treatment group whereas it was completely unobserved in the control group. The unknown compliance type of individuals in the control condition was handled as missing data via the EM algorithm.

In principle, between-school and within-school level parameters can be formally modeled taking into account compliance in the context of cluster-randomized trials (i.e., multilevel modeling). However, in practice, the number of clusters is often small, as it is in our study with 9 schools per condition. Fairly large numbers of clusters (preferably 50 or more) are necessary to yield accurate CACE estimates when taking a formal multilevel approach to account for nesting of teachers within school (Jo et al. 2008a, b). Therefore, to accommodate the nested data structure, we used the sandwich estimator in conjunction with the ML-EM mixture (TYPE = COMPLEX MIXTURE) to adjust the standard errors for the clustering of students within teachers. For the interpretation of the magnitude of the treatment effects across different models, the change in R^2 was calculated by comparing the model with the treatment variable and other covariates to the control-only model. In order to interpret the magnitude of effects across the different models, effect sizes (*ES*) were calculated by dividing the outcome difference across the two conditions by the square root of the total variance obtained from a fully unconditional model.

To facilitate a direct contrast between the CACE effects with ITT analyses, we fit identical two-class latent models across the two sets of results, which included identical covariates [i.e., baseline TOCA-R scores for the related scale, student gender, student special education status, and the students' race (dummy coded for Black and Hispanic)]. Separate models were conducted for each of the six student outcomes (i.e., academic engagement, social competence, positive peer relations, emotion regulation, inattention/hyperactivity; and aggressive-disruptive behavior) with intervention status modeled as a predictor across both the ITT and CACE analyses. Below, we report the results first for the ITT and then the CACE, with both medium and high levels of compliance, and with and without exclusion restriction assumptions for each of the six teacher-rated outcomes.

Results

ITT Estimates

We first sought to determine the effects of the intervention using a typical ITT analysis approach; these results are presented in Table 3. On average, the ITT results indicated that the students in the integrated condition were rated as displaying higher levels of social competence ($b = 0.21$, $p < .05$, effect size [*ES*] = .01) than students in the control condition, thereby suggesting a positive but small impact of the intervention for high-risk students on this outcome. None of the other ITT effects were statistically significant. In addition,

Table 3 ITT and CACE results for medium compliance

	ITT			Medium compliance								
				With ER			Without ER					
				Compliers			Compliers			Noncompliers		
	Slope	(*SE*)	*ES*	Slope	(*SE*)	*ES*	Slope	(*SE*)	*ES*	Slope	(*SE*)	*ES*
Academic engagement												
P2P v. control	0.17	0.11	0.00	0.27	0.18	0.01	0.15	0.16	0.01	0.19	0.19	0.00
PAX v. control	− 0.02	0.09	0.00	− 0.13	0.17	0.01	− 0.31	0.17	0.02	0.25	0.20	0.02
Social competence												
P2P v. control	0.21	0.10*	0.01	0.37	0.15*	0.07	0.41	0.14**	0.01	− 0.08	0.23	0.05
PAX v. control	0.04	0.09	0.00	0.05	0.22	0.00	0.16	0.18	0.02	− 0.12	0.25	− 0.01
Peer relations												
P2P v. control	0.11	0.10	0.00	0.19	0.17	0.01	0.12	0.20	0.00	0.10	0.24	0.00
PAX v. control	− 0.10	0.08	0.00	− 0.28	0.17	0.03	0.18	0.12	0.02	− 0.39	0.17	0.06
Emotion regulation												
P2P v. control	0.16	0.10	0.01	0.25	0.16	0.02	0.29	0.28	0.02	− 0.02	0.47	0.00
PAX v. control	0.04	0.10	0.00	− 0.04	0.19	0.00	0.18	0.13	0.01	− 0.14	0.18	0.00
Hyperactivity												
P2P v. control	− 0.07	0.09	0.00	− 0.15	0.19	0.02	− 0.01	0.18	0.00	− 0.14	0.25	0.01
PAX v. control	0.06	0.07	0.00	0.15	0.14	0.00	− 0.04	0.15	0.02	0.18	0.16	0.00
Aggressive-disruptive												
P2P v. control	− 0.10	0.11	0.00	− 0.21	0.21	− 0.01	− 0.34	0.19	0.02	0.17	0.17	0.03
PAX v. control	0.06	0.09	0.00	0.11	0.20	0.00	− 0.26	0.16	0.02	0.42	0.16**	0.05

Covariates included baseline TOCA-R scores for the related scale, student gender, student special education, and students' racial/ethnic group (dummy-coded variables for Black and Hispanic)

ITT intent to treat, *CACE* complier average causal effect, *SE* standard error, *ES* effect size, *P2P* integrated PATHS to PAX; PAX = PAX GBG only; ER = Exclusion restriction;

*$p < .05$, **$p < .01$

the effect sizes were less than .01 for all other outcomes for both comparisons (PATHS to PAX vs. control; and PAX GBG only vs. control).

CACE Estimates

Table 3 presents the findings for CACE analyses with the covariates of baseline scores, student gender, referral for special education, and dummy-coded variables for Black and Hispanic racial/ethnic groups among students whose teachers practiced a medium level of compliance. We also show results with the exclusion restriction (ER) assumption and without the ER assumption.

We found statistically significant effects on students' social competence for teachers who had a medium level of compliance with the treatment. Assuming the ER, students exposed to the integrated intervention were significantly more likely to be rated as having higher social competence than those in the control group ($b = .37$, $p < .05$, $ES = 0.01$). About 7% of variation in the outcome can be attributed to the treatment effect. Without the ER, teachers rated students in the integrated condition reported a 0.41 unit (on the 6-point scale) higher level of social competence than control group students, after controlling for prior social competence and demographic variables. We did not find any significant differences on the other outcomes between students exposed to the integrated and control conditions. Moreover, the only significant difference between PAX GBG only students and control students was for aggressive-disruptive behavior using the medium compliance level in classrooms of non-compliers when estimating without the exclusion restriction criterion. Specifically, the students in the PAX GBG only condition who were also in classrooms of non-compliers actually had worse levels of aggressive-disruptive behavior at posttest relative to controls (see Table 3).

As shown in CACE results in Table 4, significant effects of the integrated intervention were observed on students' academic engagement, social competence, and emotion regulation outcomes for students whose teachers demonstrated high compliance with treatment assignment. With the ER assumption, students receiving the integrated intervention were rated as displaying higher levels of academic engagement ($b = .89$, $p < .05$, $ES = 0.06$) and social competence ($b = .62$, $p < .01$, $ES = 0.17$) than students in control condition. After controlling for baseline scores and student demographics, 6% and 17% of variation in academic engagement and social competence, respectively, was explained by treatment effects. For models without the ER, teachers reported higher levels of student academic engagement ($b = .78$, $p < .05$; $ES = .04$), social competence ($b = .29$, $p < .01$; $ES = .28$), and emotion regulation ($b = .25$, $p < .05$; $ES = .02$) for those receiving the integrated interventions, as compared to the control condition students. Like the medium compliance group, we did not find significant effects of PAX GBG only intervention on any of the outcomes. In addition, there were no significant associations found for the students whose teachers were in the non-complier group.

Discussion

Given the increasing concerns regarding variation in implementation fidelity of school-based and other preventive interventions (Domitrovich et al. 2008), researchers need a range of methodological approaches that account for these issues when estimating intervention effects. CACE appears to be a promising approach for assessing impact in prevention RCTs, as it enables the inclusion of fidelity (i.e., compliance) when determining program effects. Toward that end, the primary goal of the current study was to extend our prior work on teacher outcomes by examining the impacts of the PAX GBG universal, behavioral management model, both in isolation and when integrated with the PATHS social-emotional learning curriculum, on a range of student outcomes for students with high levels of baseline aggressive-disruptive behavior.

We leveraged prior findings from a CACE study on teacher outcomes (Berg et al. 2017), which served as an initial application of this method for estimating impacts on teacher outcomes in this RCT. Other novel aspects of this application of CACE include the consideration of intervention compliance, defined here as implementation dosage at the classroom level. Similarly, we were sensitive to multiple levels of compliance (i.e., medium and high), given that dosage, used to assess compliance, was collected as a continuous variable (i.e., amount of game use) rather than dichotomous (compliant vs. non-compliant).

With regard to our primary findings, we first conducted ITT analyses with the same sample of high-risk students and with the same set of covariates in order to contrast the CACE findings accordingly. The ITT results suggested one promising effect for the integrated program, over the PAX GBG only, when compliance was not taken into consideration; specifically, the ITT analyses indicated a significant effect of the integrated PATHS to PAX condition on social competence relative to controls for students with elevated behavioral risk. While favorable, and largely consistent with some other ITT findings from this RCT on the full sample of students, the CACE findings do not signal as consistent or broad-reaching an impact of either the integrated or the PAX GBG only programs relative to controls, as when compliance is not accounted for. However, when conducting the CACE analyses, we found a number of other significant effects generally favoring the integrated PATHS to PAX intervention, particularly for students of high compliance teachers. More specifically, the CACE results indicated that

Table 4 CACE results for high compliance

	High compliance								
	With ER			Without ER					
	Compliers			Compliers			Noncompliers		
	Slope	(*SE*)	*ES*	Slope	(*SE*)	*ES*	Slope	(*SE*)	*ES*
Academic engagement									
P2P v. control	0.89	0.36*	0.06	0.78	0.38*	0.04	0.09	0.12	0.01
PAX v. control	0.03	0.61	− 0.01	0.23	0.26	− 0.02	− 0.12	0.12	0.02
Social competence									
P2P v. control	0.62	0.18**	0.17	0.29	0.10**	0.28	− 0.67	0.58	− 0.02
PAX v. control	0.11	0.30	− 0.04	0.16	0.09	0.03	− 0.50	0.42	0.03
Peer relations									
P2P v. control	0.27	0.19	0.01	0.16	0.20	0.01	0.10	0.11	0.00
PAX v. control	− 0.84	0.50	− 0.02	− 0.19	0.12	0.02	0.27	0.20	− 0.04
Emotion regulation									
P2P v. control	0.30	0.20	0.00	0.25	0.11*	0.02	− 0.70	0.72	0.12
PAX v. control	0.04	0.23	0.00	0.04	0.12	0.00	0.00	0.25	0.00
Hyperactivity									
P2P v. control	− 0.13	0.22	0.00	− 0.06	0.10	0.00	− 0.08	0.18	0.00
PAX v. control	0.48	0.47	0.01	− 0.05	0.09	0.03	0.54	0.38	0.03
Aggressive-disruptive									
P2P v. control	− 0.01	0.49	0.00	0.12	0.77	− 0.01	− 0.06	0.14	0.00
PAX v. control	0.23	0.26	0.02	0.14	0.41	0.01	0.06	0.16	0.00

Covariates included baseline TOCA-R scores for the related scale, student gender, student special education, and students' racial/ethnic group (dummy-coded variables for Black and Hispanic)

CACE complier average causal effect, *SE* standard error, *ES* effect size, *P2P* integrated PATHS to PAX, *PAX* PAX GBG only, *ER* exclusion restriction

*$p < .05$, **$p < .01$

the students with elevated aggressive-disruptive behavior who received the integrated intervention when implemented with high teacher compliance tended to demonstrate stronger effects for social competence relative to the ITT results. Moreover, two other significant effects emerged: academic engagement and emotion regulation. Interestingly, students in the medium compliance integrated group only demonstrated better social competence relative to controls, as was also demonstrated in the ITT analysis. Nevertheless, the effect sizes were all fairly small (i.e., .05 or less), with the exception of social competence which was moderate (i.e., > .20). It is notable that this was the only significant finding in the ITT analysis and may suggest that moderate effects are needed among high implementers in order to demonstrate main ITT effects. This finding may also suggest that the "medium" implementation dosage within this sample was not sufficient to positively impact students and that the "high" implementation is the minimum needed for small, but consistent, student outcomes within a one-year time frame.

Together, these findings suggest that, taking compliance into consideration, we were able to identify significant and favorable effects of the integrated program. This pattern of findings isolated a potential cut point for implementation dosage of PAX GBG that may be adequate for producing effects for students (i.e., high implementation, which was approximately 214 games played and 2110 min). On the other hand, the effects were still relatively modest and localized to the integrated program as compared to the PAX GBG only. It appears that integrated programs addressing social, emotional, and behavioral outcomes may yield more favorable and far-reaching student outcomes than narrower programs only focused on one area (e.g., behavior; see Ialongo et al. 2019).

It is also important to note that no iatrogenic effects of the program were observed in the current study, except for aggressive-disruptive behavior among the non-compliers for the PAX GBG only condition, in the model without ER. This finding suggests that in order to improve aggressive-disruptive behavior, teachers may need to ensure high compliance, particularly when implementing the PAX GBG only, as low dosage implementation may actually exacerbate the students' aggressive-disruptive behaviors. In contrast, prior findings reported by Berg et al. (2017) suggested that teachers who were highly compliant in the integrated condition actually experienced elevated levels of burnout and/or emotional exhaustion. Although the teacher CACE analysis outcomes suggested that implementing more program content may have placed additional burden on, and thus burnout for teachers, the current pattern of student results did not suggest that any such negative effects emerged for their these students under any of the conditions tested, or for any of the outcomes. This could suggest either that burnout did not translate into an increase in negative teacher perceptions or that negative perceptions did not impact the students within the time period assessed.

Limitations and Future Directions

The sample size of classrooms was relatively small, and thus became even smaller when split into implementation conditions and compliance groups. Estimation of program impacts using the high implementation cut point yielded an especially small sample size and thus may have resulted in unstable estimates. In addition, despite the fact that schools and not teachers were randomized, the small sample of schools in each condition prevented us from employing a multilevel modeling approach to account for the clustering within schools. The sample was also reduced somewhat due to missing data at either the student or teacher level; due to the complexity of the analyses, multiple imputation was not feasible. The multilevel mixture modeling using the EM estimation approach is computationally demanding, particularly when the number of clusters is small (Jo et al. 2008a, b); therefore, we were not able to accommodate the clustering of teachers in schools. Consistent with findings from prior studies of GBG (e.g., Kellam et al. 1998), we focused our analyses on the at-risk subsample of students who at baseline were at elevated risk on aggressive-disruptive behavior.

Our interpretation of findings is also limited by our indicator of implementation, which unlike prior CACE applications did not allow for the definition of "full implementation" using a theoretically-driven cutpoint of dosage (e.g., Stuart et al. 2008) or a clear binary indicator (e.g., Connell et al. 2007; Cowen 2008). While the primary focus of this and the Berg et al. (2017) studies was outcomes for students and teachers, respectively, both also served to define cut points on a continuum of medium to high dosage based on their associations with outcomes and may be a useful guide for approximating high dosage in future studies of GBG. A potential limitation of the current study was that we focused only on one type of compliance, which was dosage, as our prior research suggested relatively limited variability in implementation quality, particularly in contrast to the level of variability in the dosage indicator (Domitrovich et al. 2015). Specifically, there appears to be a bit of a ceiling effect on the implementation quality rating (as assessed by trained observers); thus, there was too little variability in the quality indicator to yield a meaningful contrast in compliance levels.

In addition, CACE estimation relies on a set of assumptions, which were largely met in this study. On the other hand, our models assumed additive effects of dosage, which relaxed the exclusion restriction and therefore were more likely to suffer from a violation of normality. To address

these possible violations, we conducted the CACE estimation with two different cut points and with and without the assumption of the ER (Jo 2002) as a form of sensitivity analyses. These analyses provided greater support for our findings, which were generally consistent across outcomes, but also resulted in a large number of statistical tests conducted. Finally, we assessed the student outcomes through teacher report on the TOCA-R; as such the same teachers who implemented the intervention also delivered the intervention, and thus were aware of their own intervention status. This source of shared measurement may have influenced their ratings on the TOCA-R (Pas and Bradshaw 2014).

With regard to future research directions, one might explore potential factors that predict teachers' compliance with implementation. Prior research on these data identified that initial impressions of the game and its fit with the needs of the classroom (see Domitrovich et al. 2015), teacher emotional exhaustion, and the level of student aggressive behavior in the classroom were associated with implementation levels (Musci et al. 2019). Additional research is needed to better understand if these or other variables (such as years of experience teaching, proportion of students with high need in their classroom, teacher efficacy, or school climate) predict compliance or are related to implementation of preventive interventions more generally (see Domitrovich et al. 2008). Such analyses could also explore the extent to which these predictors of compliance varied by intervention condition.

Conclusions and Implications

Keeping in mind the limitations outlined above, the findings generally demonstrated some promise that when implemented with high dosage, the PATHS to PAX program results in positive outcomes for students at elevated behavioral risk, relative to controls. The effects were less salient for the PAX GBG only model; this was true in both the ITT and CACE analyses. Yet it appears that compliance (i.e., dosage) mattered, such that more consistent student effects were detected for students who received high dosage of the GBG within the integrated model; generally speaking, this was not the case for the PAX GBG only program. This study has implications for preventive intervention research, as it provides further evidence of the importance of considering implementation compliance when estimating program effects, even in the context of efficacy studies. It also establishes some benchmarks for adequate fidelity within the PAX GBG literature and has implications for the value of integrated prevention approaches that target multiple student skills.

There are also important implications of this work for school administrators, as they play a critical role in setting expectations for teachers' high compliance with implementation (Domitrovich et al. 2008). Moreover, policies should also reflect a stronger attention not just to the adoption of evidence-based practices, but also specify expectations for high compliance. Additional supports and funding are needed to encourage deeper investment in implementation supports at the school and classroom levels, including a strong commitment to providing coaching, on-going professional development, monitoring for compliance, and incentives for high fidelity implementation. Such investments are critical to ensuring that the intended outcomes of prevention programs are realized through high fidelity implementation.

Acknowledgements This research was supported in part by grants from the Institute of Education [R305A080326; R305A130060] and the National Institute of Mental Health [P30 MH08643]. The authors would like to thank Celene Domitrovich for her contributions to the project and Booil Jo for consultation on the analyses.

References

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

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. (2003). Principal stratification approach to broken randomized experiments. *Journal of American Statistical Association, 98*, 299–323. https://doi.org/10.1198/016214503000071.

Berg, J., Bradshaw, C. P., Jo, B., & Ialongo, N. S. (2017). Using complier average causal effect estimation to determine the impacts of the Good Behavior Game preventive intervention on teacher implementers. *Administration and Policy in Mental Health and Mental Health Services Research, 44*, 558–571. https://doi.org/10.1007/s10488-016-0738-1.

Bradshaw, C. P., & Kush, J. M. (2019). Teacher observation of classroom adaptation-checklist: Measuring children's social, emotional, and behavioral functioning. *Children & Schools, 42*(1), 29–40. https://doi.org/10.1093/cs/cdz022.

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2015). Examining variation in the impact of school-wide positive behavioral interventions and supports: Findings from a randomized controlled effectiveness trial. *Journal of Educational Psychology, 107*(2), 546–557.

Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology, 101*(4), 926–937. https://doi.org/10.1037/a0016586.

Chen, H. T. (1998). Theory-driven evaluations. *Advances in Educational Productivity, 7*, 15–34.

Conduct Disorder Problems Research Group. (1999). Initial impact of the Fast Track prevention trial for conduct problems II: Classroom effects. *Journal of Consulting and Clinical Psychology, 67*, 648–657.

Connell, A. M., Dishion, M. Y., Yasui, M., & Kavanagh, K. (2007). An adaptive approach to family intervention: Linking engagement in family-centered intervention to reductions in adolescent behavior. *Journal of Consulting and Clinical Psychology, 75*, 568–579. https://doi.org/10.1037/0022-006X.75.4.568.

Cowen, J. M. (2008). School choice as a latent variable: Estimating the "complier average causal effect" of vouchers in Charlotte. *The Policy Studies Journal, 36*, 301–315.

Domitrovich, C., Bradshaw, C., Greenberg, M., Embry, D., Poduska, J., & Ialongo, N. (2010). Integrated models of school-based prevention: Logic and theory. *Psychology in the Schools, 47*, 71–88.

Domitrovich, C., Bradshaw, C. P., Berg, J., Pas, E., Becker, K., Musci, R., et al. (2016). How do school-based prevention programs impact teachers? Findings from a randomized trial of an integrated classroom management and social-emotional program. *Prevention Science, 17*, 325–337.

Domitrovich, C., Pas, E., Bradshaw, C. P., Becker, K., Keperling, J. P., Embry, D., et al. (2015). Individual and school organizational factors that influence implementation of the Pax Good Behavior Game intervention. *Prevention Science, 6*(8), 1064–1074.

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K. E., Buckley, J. A., Olin, S., et al. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion, 1*, 6–28.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.

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

Embry, D., Staatemeier, G., Richardson, C., Lauger, K., & Mitich, J. (2003). *The PAX good behavior game* (1st ed.). Center City, MN: Hazelden.

Fixsen, D., Naoom, S., Blasé, K., Friedman, R., & Wallace, F. (2005). Implementation research: A synthesis of the literature. Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute. The National Implementation Research Network (FMHI Publication #231).

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

Greenberg, M., & Kusché, C. (2006). Building social and emotional competence: the PATHS Curriculum. In S. R. Jimerson & M. J. Furlong (Eds.), *Handbook of school violence and school safety: From research to practice* (pp. 395–412). Mahwah, NJ: Erlbaum.

Greenberg, M., Kusché, C., Cook, E., & Quamma, J. (1995). Promoting emotional competence in school-aged children: The effects of the PATHS curriculum. *Development and Psychopathology, 7*, 117–136.

Greenberg, M. T., Kusché, C. A., & Conduct Problems Prevention Research Group. (2011). *Grade level PATHS (Grades 3–5)*. South Deerfield, MA: Channing-Bete Co.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–970.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*, 88–110.

Ialongo, N., Poduska, J., Werthamer, L., & Kellam, S. (2001). The distal impact of two first-grade preventive interventions on conduct problems and disorder in early adolescence. *Journal of Emotional and Behavioral Disorders, 9*, 146–160.

Ialongo, N., Wethamer, L., Kellam, S., Brown, C., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and anti-social behavior. *American Journal of Community Psychology, 27*, 599–641.

Ialongo, N. S., Domitrovich, C. E., Embry, D., Greenberg, M., Lawson, A., Becker, C., et al. (2019). A randomized controlled trial of the combination of two school-based universal preventive Interventions. *Developmental Psychology, 56*(6), 1313–1325. https://doi.org/10.1037/dev0000715.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27*, 385–409.

Jo, B., Asparouhov, T., & Muthén, B. O. (2008a). Intention-to-treat analysis in cluster randomized trials with noncompliance. *Statistics in Medicine, 27*, 5565–5577. https://doi.org/10.1002/sim.3370.

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008b). Cluster randomized trials with treatment noncompliance. *Psychological Methods, 13*, 1–18.

Kellam, S., Brown, C. H., Poduska, J., Ialongo, N., Wang, W., Toyinbo, P., et al. (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, 95S*, S5–S28.

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

Kellam, S. G., Wang, W., Mackenzie, A. C. L., Brown, C. H., Ompad, D. C., Or, F., et al. (2014). The impact of the Good Behavior Game, a universal classroom-based preventive intervention in first and second grades, on high-risk sexual behaviors and drug abuse and dependence disorders into young adulthood. *Prevention Science, 15*, 6–18. https://doi.org/10.1007/s11121-012-0296-z.

Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher Observation of Classroom Adaptation-Checklist (TOCA-C): Development and factor structure. *Measurement and Evaluation in Counseling and Development, 42*, 15–30. https://doi.org/10.1177/0748175609333560.

Kusché, C. A., Greenberg, M. T., & Conduct Problems Prevention Research Group. (2011). *Grade level PATHS (Grades 1–2)*. South Deerfield, MA: Channing-Bete Co.

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

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

Musci, R., Pas, E. T., Bettencourt, A., Masyn, K., Ialongo, N. S., & Bradshaw, C. P. (2019). How does collective student behavior and other classroom contextual factors relate to implementation of an evidence-based intervention? A multilevel SEM. *Development and Psychopathology, 31*, 1827–1835. https://doi.org/10.1017/s095457941900097x.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*, 33–84.

Pas, E., & Bradshaw, C. P. (2014). What affects teacher ratings of student behaviors? The potential influence of teachers' perceptions of the school environment and experiences. *Prevention Science, 15*, 940–950. https://doi.org/10.1007/s11121-013-0432-4.

Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014-4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, national Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from https://ies.ed.gov/ncee/edlabs.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate: Causal inference in the face of interference. *Journal of the American Statistical Association, 101*, 1398–1407.

Stanger, C., Ryan, S. R., Fu, H., & Budney, A. J. (2011). Parent training plus contingency management for substance abuse families: A Complier Average Causal Effects (CACE) analysis. *Drug and Alcohol Dependence, 118*, 119–126. https://doi.org/10.1016/j.drugalcdep.2011.03.007.

Stuart, E. A., Perry, D. F., Le, H., & Ialongo, N. S. (2008). Estimating intervention effects of prevention programs: Accounting for noncompliance. *Prevention Science, 9*, 288–298. https://doi.org/10.1007/s11121-008-0104-y.

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