Journal of School Psychology 78 (2020) 75–95

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Randomized trial testing the integration of the Good Behavior Game and MyTeachingPartner™: The moderating role of distress among new teachers on student outcomes☆

Patrick Tolan[a,*], Lauren Molloy Elreda[a], Catherine P. Bradshaw[a], Jason T. Downer[a], Nicholas Ialongo[b]

[a] University of Virginia, United States of America
[b] Johns Hopkins University, United States of America

ARTICLE INFO

Action Editor: Keith Herman

Keywords:
Teacher training
Stress
Classroom management
Randomized controlled trial
New Teachers
Intervention

ABSTRACT

A growing body of research documents the effectiveness of classroom management programs on a range of student outcomes, yet few early-career teachers receive training on these practices prior to entering the classroom. Moreover, few studies have attended to how variations in teacher distress or level of classroom misbehavior affects training benefits. This study reports findings from a randomized trial of a teacher training program that combined two evidence-based programs (Good Behavior Game [GBG] and MyTeachingPartner™ [MTP]) to determine their impact on novice teachers and their students. In addition, the current study reports findings on moderated impacts by initial teacher distress as well as the overall classroom level of misbehavior. The sample included 188 early-career teachers (grades K-3) in their first three years of teaching from three large, urban school districts. Analyses indicated that intervention had no main effects, but yielded moderated impact depending on the combination of the baseline levels of classroom disruptive behavior and teacher distress; it appears that the program impacts were greatest in the highest risk circumstance (i.e., high teacher stress and elevated challenging student behaviors). For those classrooms, those assigned to intervention evidence improved behavior and student achievement compared to control counterparts by the spring of the training year, relative to the fall baseline (d = 0.18–0.70 depending on outcome). This study is significant in that it highlights effects during a critical window of training and coaching for early career teachers and the need to consider teacher and classroom contextual factors that may moderate professional development efforts.

1. Introduction

Considerable evidence from longitudinal studies shows that child aggressive/disruptive behavior in early elementary school predicts later mental health and behavioral problems, as well as lower educational and occupational attainment (Dubow, Boxer, & Huesmann, 2009; Kellam et al., 2008; Kellam, Rebok, Ialongo, & Mayer, 1994; Thomas, Bierman, and The Conduct Problems

☆ This article is part of the special issue 'Advances in Understanding and Intervening in Teacher Stress and Coping; Edited by Dr. Keith Herman, Dr. Wendy Reinke, and Ms. Colleen Eddy.

* Corresponding author at: Charles S. Robb Professor of Education and Human Development, Youth-Nex Center and Curry School of Education and Human Development, University of Virginia, United States of America.
E-mail address: pht6t@virginia.edu (P. Tolan).

https://doi.org/10.1016/j.jsp.2019.12.002
Received 21 December 2018; Received in revised form 21 October 2019; Accepted 8 December 2019
Available online 08 February 2020

Prevention Research Group, 2006). Additionally, the level of classroom disruptive behavior has been shown to affect subsequent child aggression controlling for individual behavior (Thomas et al., 2006). Accordingly, a substantial portion of prevention efforts have focused on classroom behavior management efforts through teacher training.

Multiple prevention trials have demonstrated that programs designed to improve classroom management practices can have a direct influence on disruptive behavior, self-control, and academic failure, which are risk factors for long-term problems in multiple domains of functioning (Bradshaw, Zmuda, Kellam, & Ialongo, 2009b; Greenberg et al., 2003; Petras, Masyn, & Ialongo, 2011). Classroom management practices can also indirectly affect child outcomes through influence on student engagement (Oliver & Reschly, 2007), and teacher satisfaction and motivation (Harmsen, Lorenz, Maulana, & van Veen, 2018; Rimm-Kaufman & Hamre, 2010). Together, the extant research suggests a moderate-sized effect of teacher training in classroom management on students' subsequent risk for mental health, substance use, academic performance, and other important outcomes. Moreover, these findings suggest better long-term outcomes when the training targets students early in the early elementary grades (Ialongo et al., 2006; Oliver, Wehby, & Daniel, 2011).

The extensive literature on the benefits of teacher training in classroom behavior management supports an emphasis on teacher training to affect classroom behavior during early schooling. In addition to the empirical evidence that such training can have lasting benefits for those students who are in the classroom during the prevention training period, this approach is also justified by the supposition that by training teachers, benefits can extend to multiple cohorts of students who *subsequently* enter their classrooms. Unfortunately, too many newly degreed teachers have not had direct training in classroom management and fewer have had direct training using empirically proven methods (Kane, Rockoff, & Staiger, 2006). Moreover, early career teachers, in particular, report concerns about lacking effective means to handle the disruptive behavior of students (Browers & Tomic, 2000; Greenberg, Brown, & Abenavoli, 2016). For these reasons understanding how classroom management training might be most effectively focused with novice teachers is important.

Another important but minimally studied consideration is how teacher and classroom characteristics may contribute to variation in classroom management practices and effects on students (Greenberg & Abenavoli, 2017). There is a growing number of studies of preventive interventions that have reported variation in program effects based on setting and baseline conditions (Bradshaw, Waasdorp, & Leaf, 2015; Domitrovich et al., 2015; Ialongo, Werthamer, Brown, Kellam, & Wai, 1999; Kellam et al., 1994). In fact, the impact of teacher training and other such preventive interventions are rarely consistent across all settings, students, or implementers (Domitrovich et al., 2015). Two potential moderators of impact of training for new teachers are 1) the distress level the teacher experiences as they start teaching (Greenberg et al., 2016) and 2) how much misbehavior they face in their initial classroom experience (American Federation of Teachers, 2015; Musci, Pas, Bettencourt, Masyn, Ialongo, & Bradshaw, 2019).

Teacher stress and attitudes reflecting burnout can undermine investment in teaching quality and intrinsic career value (Ema Ushioda and Dörnyei, 2012). A recent survey of over 30,000 teachers noted that nearly three-quarters experienced their work as stressful (American Federation of Teachers, 2015), and several studies indicate that stress may be related to the quality of teachers' practice and student outcomes (Arens & Morin, 2016; Hoglund, Klingle, & Hosan, 2015). Teachers vary in readiness for and ability to cope with demands of classroom management (Greenberg et al., 2016). Some teachers enter classrooms feeling prepared and experience challenges as energizing opportunities for growth, whereas others may experience these same conditions as overwhelming, disillusioning, and emotionally draining. Therefore, it is important to understand the extent to which the impacts of such trainings vary as a function of teachers' experience and perception of stress and related emotional distress.

Similarly, classrooms vary considerably in how challenging and disruptive the student behavior is at the outset of the school year (Ma & Willms, 2004); teachers are presented with classrooms of students that vary considerably in how compliant, on task, and disruptive they are (Jennings & Greenberg, 2009). Although there has been little examination of how classroom baseline features affect student outcomes, findings from several prevention trials suggest that interventions in classrooms may be most influential for the most problematic students (Ialongo et al., 1999; Kellam et al., 1994). Therefore, it is important to explore for variation in program effects as a function of baseline classroom disruptive behavior level, as it may be systematically associated with differences in program outcomes which in turn inform the targeting for limited training resources (Domitrovich et al., 2009; Musci et al., 2019).

To address these gaps, the current study was designed to test, through a randomized controlled trial design, the combined impact of two promising evidence-based classroom management and teacher support programs. Specifically, we integrated training and support in the implementation of the Good Behavior Game (Ialongo et al., 1999) with MyTeachingPartner™ (Pianta, La Paro, Payne, Cox, & Bradley, 2002), to determine their combined impact on new elementary school teachers and their students. In addition, we examined moderation of the effects by an important teacher characteristic (i.e., level of distress about teaching at baseline) and by the setting in which initial teaching occurs (i.e., classroom-level student disruptive behavior level at baseline). This line of research has important implications not only for the prevention of behavioral and academic problems in students, but for supporting new teachers who are entering the field of education. As further illustrated below, this study also builds on prior research documenting the needs of novice teachers.

1.1. Focusing on new teachers

Few studies to date have considered issues related to teacher experience when providing training in evidence-based prevention programs. However, it is quite likely that motivation, benefits, and likelihood of continuing practices vary as a function of length of time teaching and the extent to which practices have become engrained. Just as preventive benefits for youth are predicated on the principle of early intervention, when influences on later behavior are malleable, it seems likely that teachers may be more amenable to training and professional development earlier in their career, as compared to mid-career or later. In fact, reviews show that more

teacher experience is related to better student achievement, with the greatest impact occurring during the first five years of teaching (Clotfelter, Ladd, & Vigdor, 2007). Additionally, novice teachers typically report more optimism that training will improve their effectiveness. Overall, novice teachers may be more open to innovations and receptive to feedback, as they are still developing effective practices (Browers & Tomic, 2000).

Also, it seems likely that more experienced teachers, with habits of practice that are well established, are less likely to be enthusiastic about training in basic classroom management and more prone to sustaining established approaches (Harmsen et al., 2018). Focusing on newer teachers, who are just forming practice habits and likely have greater motivation to attain skills that enable competent classroom management, seems advisable. Motivation to learn, receptivity to guidance, and likelihood to sustain trained practices all are likely to be stronger for those just entering teaching than those with many years of experience. Enabling effective teaching rather than undoing problematic engrained habits seems more likely to maximize impact. For these reasons and due to the absence of consideration of teacher experience levels in most other tests of teacher training programs to affect classroom management, we focused on teachers shortly after they have transitioned from preservice training into classroom teaching.

1.2. Professional development needs of new teachers

Studies of effective teaching have identified three major gaps and areas of need for professional development of new teachers: classroom behavioral management (Ialongo et al., 1999; Kellam et al., 1994), a positive and engaged relationship around learning between the teacher and students (Evertson & Weinstein, 2006; Pianta, Belsky, Houts, & Morrison, 2007), and organization of the classroom to fit the learning tasks and to optimize opportunity for good student behavior (Oliver & Reschly, 2007). For example, a meta-analysis examining aspects of teacher quality in relation to achievement found significant and substantial associations between each of these components and disruptive behavior in classrooms: *quality student-teacher interactions* (average effect size $d = -0.87$), *discipline/incentive methods used* (average effect size of $d = -0.91$) and *establishment and reliance on clear and understandable rules and behavioral expectations* (average effect size of $d = -0.76$) (Marzano, Gaddy, & Dean, 2000). Use of these effective methods also related to an average effect size of $d = 0.62$ increase in student engagement and $d = 0.52$ gain in achievement (approximately a 20 percentile point increase).

Prior research has identified specific professional development models which have demonstrated positive effects on one or more of these focal areas. One such program is the Good Behavior Game [GBG], which targets classroom behavior management through group contingencies (Bradshaw et al., 2009b; Bradshaw, Koth, Thornton, & Leaf, 2009a; Ialongo et al., 1999; Kellam et al., 1994). In multiple randomized controlled trials (RCTs) the GBG has been shown to reduce student disruptive behavior and aggression, with the most consistent and lasting effects for the students with elevated rates of such behavior prior to implementation (Flower, McKenna, Bunuan, Muething, & Vega, 2014). In most tests as a universal preventive effort in elementary school settings the effects tend to be small to moderate, but lasting and with long term impact on for important problems such as drug use and mental illness rates (Kellam et al., 2008). For example, in a prior trial of the approach utilized here, effect sizes for reductions relative to controls in aggressive/disruptive and off-task behavior by the end of first grade based on independent observations (Brown, 1993), teacher ratings, and peer nominations (Dolan et al., 1993) were generally about Cohen's d of ~ 0.2. In another trial, Bradshaw et al. (2009a, 2009b) documented long-term benefits; the GBG when combined with an academic intervention was associated with higher achievement scores, and greater odds of high school completion, college attendance, and not having special education services as compared to students in the control condition.

Another such program is MyTeachingPartner™ [MTP], which has a particular focus on promoting relationship-based engagement of students and supportive, enriching interactions among teachers and students (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Pianta et al., 2002). Classroom management is a key emphasis in MTP, but there is also attention paid to enhancing teachers' provision of emotional and instructional support. Prior RCT evaluations of MTP have shown improvement in teacher use of high quality teacher-student interactions ($ds = 0.43$–0.97 for preschool; Downer et al., 2018; Pianta, Mashburn, Downer, Hamre, & Justice, 2008, $ds = 0.31$–0.50 for middle school; Allen et al., 2011), as well as academic and social outcomes in preschool and secondary school settings (Allen et al., 2011; Allen, Hafen, Gregory, Mikami, & Pianta, 2015; Hamre et al., 2012a; Hamre, Pianta, Mashburn, & Downer, 2012b; Mashburn, Downer, Hamre, Justice, & Pianta, 2010; Mikami, Gregory, Allen, Pianta, & Lun, 2011).

Thus, both GBG and MTP have shown significant beneficial impact on student behavior and achievement through change in teacher practices, but each emphasizes a different aspect of classroom management. GBG is a group contingency classroom behavior management and instructional support approach that rewards positive group, as opposed to individual, behavior (Kellam et al., 1994). The team-based nature of the game allows teachers to take advantage of peer pressure in managing student behavior. In contrast, MTP emphasizes how teachers promote high-quality interactions with and between their students to maximize engagement in learning.

Although the specific elements of these two models are described in more detail in the methods section, briefly, GBG leverages social learning principles of peer and teacher reinforcement, which numerous studies have demonstrated is linked with reductions in disruptive behavior and improvements in academic performance (Bradshaw et al., 2009a, 2009b; Ialongo et al., 1999; Kellam et al., 1994). MTP is a web-mediated, video-based coaching approach based on the hypothesis that teachers require extensive opportunities for: (a) observation of effective interactions via analysis of multiple video examples, (b) identifying effective interactive responses to students' cues, and (c) ongoing individualized feedback and problem-solving analysis of their own interactions (e.g., Allen et al., 2011; Pianta et al., 2007). Toward this end, these two programs incorporate complementary elements which have been identified as important for classroom organization to optimize learning (Embry, Staatemeier, Richardson, Lauger, & Mitich, 2003; Evertson & Weinstein, 2006; Tomlinson & Imbeau, 2010). In addition, as designed, GBG and MTP utilize an intensive training workshop followed

by ongoing coaching of teachers for aiding implementation fidelity and competence. The complementary emphasis on important components to effective classroom management suggested these could be integrated into a comprehensive training program that addresses the simultaneous challenges of classroom management. When combined, the training can enable skill-building through coaching support that moves teachers from how one sets up the classroom and engages students at the start, to how one helps organize and motivate the group to achieve over the course of the year (Domitrovich et al., 2009). Whereas MTP focuses on overall classroom management through quality of teacher relationships with students around learning, GBG provides a specific framework targeting promotion of on-task behavior and decreases in student disruptive behavior. By combining these two approaches, we aimed to provide opportunities for improvement in classroom management through skill attainment within an overall strategy of student engagement. More specifically, by integrating these two programs together, new teachers could be provided skills and support for competency in each of the three essential components of effective classroom management (Sugai, Horner, & Gresham, 2001). Moreover, as most training of new or pre-service teachers emphasizes subject matter content with only incidental attention to and general commentary about behavior management, systematic training and coaching in these components could be used to meet often-unmet needs of new teachers for professional development related to classroom management (Neuman & Cunningham, 2009).

1.3. Overview of the current study

The current study was designed to test the combined impacts of these two evidence-based classroom management and student engagement professional development models. Specifically, we combined and integrated the training and professional development provided through these two programs, leveraging the proximal focus of the GBG on the classroom setting, and the tailored coaching and supports provided through the MTP model. We focused on novice teachers, which we defined as those in the first through third year of teaching, and those working in elementary schools (K-3), given the developmental sensitivity of this age range for optimizing impacts for students.

In addition, we included measurement of two factors present at baseline to explore the possible moderating effects of the teacher distress about teaching and how challenging the classroom to which the teacher was assigned (classroom level of student disruptive behavior). These emphases bring into this efficacy test consideration of important features that might relate to variation in effects (Domitrovich et al., 2015; Greenberg & Abenavoli, 2017).

Our specific hypotheses were that 1) classroom levels of disruptive behavior, on-task behavior, compliance with teacher directives, and academic achievement (reading and math) would improve significantly compared to no-intervention control counterparts and 2) initial levels of student disruptive behavior at the classroom level and teacher distress would moderate the extent of difference found between conditions. We expected effect sizes comparable to prior RCTs of the individual programs, although given these have not been combined before nor applied with only new teachers, we did not make specific hypotheses about effect sizes.

2. Method

2.1. Research design

We used a randomized trial design, in which students clustered within classrooms/teachers to evaluate the combined impacts of GBG and MTP (referred to as GBG + MTP) on student achievement and behavior (Murray, Varnell, & Blitstein, 2004). Specifically, teachers were randomly assigned to either receive training and coaching in the combined GBG + MTP intervention or to serve as controls (business as usual), with measurements taken on the teachers and students nested within their classrooms to assess the impact of the intervention. The current study focuses on student outcomes for those in the participating teachers' classroom *during* the training, coaching, and implementation year relative to control teachers; specifically, we contrast the fall baseline data with a spring posttest, following a single academic year of coaching supports. Notably, all behavioral data are based on observations at the classroom level; for consistency and because we are interested in classroom level effects not variation within a given classroom, we aggregated the student achievement data up to the classroom level, rendering a single level (teacher) analysis appropriate for all outcomes.

2.2. Sample

2.2.1. Teachers

Three consecutive cohorts of new teachers (with 3 years or less of prior teaching experience), who were hired by one of three participating public-school districts to full-time K-3 classroom instruction in the ensuing year of training, were recruited into the project. To limit heterogeneity in teaching demands, eligible teachers included those in early grades (K-3) and excluded Teach for America engaged teachers, given their variation in background from typical teachers. All participants were teachers with at least a bachelor's degree in education. The three school systems all tended to serve relatively impoverished communities, although they varied from a major urban district to one serving a smaller city and one serving the suburban areas surrounding the major city. Teachers were recruited from among all new teachers in each district each year, and thus were distributed across multiple schools. Therefore, in most instances schools had only a single teacher participant in each year of recruitment, with a maximum of eight teacher participants from any one school over the course of the trial. Fig. 1 provides a CONSORT chart of recruitment and retention of teachers.

Our power analyses indicated that we should aim to enroll 252 teachers for randomization to detect expectable main effect sizes

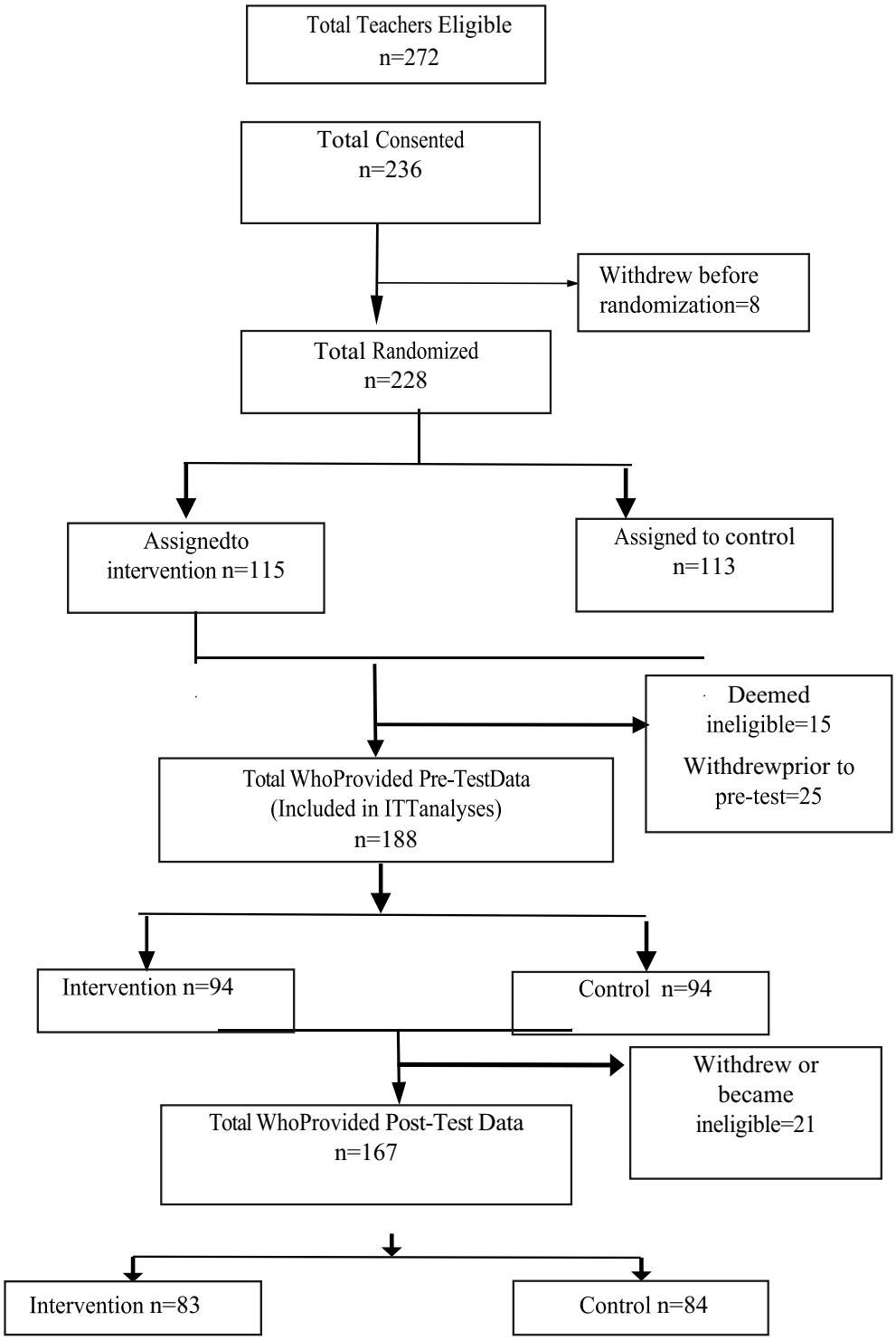


Fig. 1. CONSORT Chart for MTP + GBG Trial.

down to $d = 0.15$ with 90% probability assuming 10% attrition over the first year. With the proportion of variance in outcomes ultimately accounted for by baseline covariates ranging from 0.31 to 0.39 for the three behavioral outcomes, and being 0.98 for reading and math achievement, the resulting sample of 188 provided power (80% probability) to detect effects as low as 0.32 for the behavioral outcomes, and effects as small as 0.06 for reading and math achievement. As summarized above, these effect sizes are within the range of what has been detected for post-test effects among prior trials of these two training programs. Two-hundred and

Table 1
Comparisons of teacher characteristic by condition.

	Sample included in ITT (intent to treat) analyses		Frequencies			Frequencies		
	Frequencies	Percentage	Control group	Intervention group	Chi square test	Dropped before post-test	Completed post-test	Chi square test
Teacher gender								
Female	176	94.0%	89	87		20	156	
Male	12	6.0%	5	7	n.s.	1	11	n.s.
Teacher race								
White	142	75.5%	67	75		17	125	
African American	42	22.3%	26	16		4	38	
Hispanic/Latinx	1	0.5%	1	0		0	1	
Other or mixed race/ ethnicity	3	1.6%	0	3	n.s.	0	3	n.s.
Year of teaching								
First year	109	58.0%	53	56		11	98	
Second year	42	22.3%	22	20		3	39	
Third year	37	19.7%	19	18	n.s.	7	30	n.s.
Degree								
Bachelor's	145	77.1%	72	73		16	129	
Master's and/or Post-Master's certificate	37	19.7%	18	19	n.s.	1	36	n.s.
Missing	6	3.2%						
Grade level								
Kindergarten	44	23.4%	23	21		4	40	
1st grade	55	29.3%	27	28		8	47	
2nd grade	42	22.3%	18	24		5	37	
3rd grade	45	23.9%	25	20		4	41	
Multiple grades	2	1.1%	1	1	n.s.	0	2	n.s.

seventy-two teachers were recruited, of which 236 teachers consented to participate. Of those, 15 were deemed ineligible due to being assigned to an ineligible classroom (i.e., not K-3rd, special education classroom, resource class), not being permitted to attend training (based on principal decision), having already been trained in the GBG or MTP, leaving the participating districts, or leaving the teacher profession all together. Of those eligible, 25 left the project prior to baseline data collection, and an additional 8 withdrew before randomization, yielding the final sample of 188 teachers recruited from 72 schools (*Mdn* number of teachers per school across three cohorts = 2, *Range* = 1–13 teachers) included in our intent to treat analyses. Cohort 1 consisted of 66 teachers from 34 schools, Cohort 2 consisted of 51 teachers from 30 schools, and Cohort 3 consisted of 81 teachers from 36 schools. The majority of teachers were female (94%) and White (76%), and 23% had a master's degree, with the remainder having no higher than a bachelor's degree (77%). Teachers ranged in years of experience from 1 to 3 years, with the majority of teachers being in their first year of teaching (58%). In addition, teachers were approximately evenly distributed across the K-3 grade levels. Attrition during the study year was low, with 11% (10 from control, 11 from intervention) discontinuing participation before the post-test. We applied Full Information Maximum Likelihood to address missing data at the student and classroom levels.

Tables 1 and 2 provide summaries of teacher characteristics and baseline scores and ranges on the variables included in analyses. Tables 1 and 2 also show comparisons made of 1) participants who completed the post-test to those who dropped out before post-test and 2) participants in the control versus the treatment groups, in terms of their data at baseline for the outcome, control, and moderation variables. The control and treatment groups did not significantly differ at baseline on any of the outcomes, moderators, or control variables. In the comparisons of the group who dropped out before post-test to those who participated in the post-test, we found only one statistically significant difference: those who dropped out before post-test were more likely to be from a school with a larger percentage African American student body. This variable was entered as a control in the models run to test effects. None of the other control, moderator, or outcome variables differed between these two groups at baseline. Similarly, when baseline comparisons were conducted between subgroups formed in the moderation analyses, no significant differences were found by condition on outcomes, moderators, or control variables at baseline.

2.2.2. Students

For each teacher's classroom, 8–10 students were randomly selected for collection of student achievement data at baseline and at the end of the intervention year, and for teacher reports of students' demographic characteristics. Students were selected based on random number generation between 1 and 10 to identify selection within gender from alphabetical class lists (4–5 each; e.g., 1st, 2nd, 4th, 6th, 10th male and female students) by giving teachers a set of random numbers to correspond to order on class lists, with instructions to rank the corresponding female and male students (e.g., told to rate 1st, 4th, 6th, 8th, 10th, male and females). Classroom observations were conducted of the class as a whole, with coding and ratings of classroom level behavior patterns. In terms of the ethnicities of those students with baseline data, 61.6% were African American, 17.3% were Hispanic or Latino, 11.0% were

Table 2
Comparison of outcome, control and moderator variables by condition and attrition.

	Sample included in ITT analyses				Mean (SD)			Mean (SD)		
	Mean	Standard deviation	Min	Max	Control	Intervention	*t*-Test	Dropped before post-test	Completed post-test	*t*-Test
Post-test scores on outcome variables										
Student socially disruptive behavior	0.49	0.28	0.06	1.56	0.48 (0.30)	0.49 (0.26)	n.s.			
Student off-task behavior	2.04	1.69	0.00	11.50	2.08 (1.95)	1.99 (1.39)	n.s.			
Student compliance	2.73	0.56	0.95	3.92	2.76 (0.61)	2.69 (0.51)	n.s.			
WJ reading composite	2.39	0.88	0.74	4.53	2.38 (0.89)	2.41 (0.89)	n.s.			
WJ math composite	2.34	1.04	0.49	4.54	2.32 (1.03)	2.37 (1.06)	n.s.			
Baseline scores on outcome variables										
Student socially disruptive behavior	0.55	0.34	0.08	1.67	0.54 (0.34)	0.56 (0.34)	n.s.	0.49 (0.31)	0.55 (0.34)	n.s.
Student off-task behavior	1.90	1.25	0.00	6.88	1.92 (1.30)	1.88 (1.22)	n.s.	1.71 (1.25)	1.91 (1.26)	n.s.
Student compliance	2.73	0.58	1.10	3.92	2.75 (0.59)	2.71 (0.58)	n.s.	2.73 (0.57)	2.73 (0.58)	n.s.
WJ reading composite	2.12	0.92	0.50	4.43	2.12 (0.90)	2.13 (0.94)	n.s.			
WJ math composite	2.05	1.01	0.32	4.39	2.01 (0.99)	2.09 (1.06)	n.s.			
Moderators										
Student socially disruptive behavior	0.55	0.34	0.08	1.67	0.54 (0.34)	0.56 (0.34)	n.s.	0.49 (0.31)	0.55 (0.34)	n.s.
Teacher distress	2.61	0.60	1.00	3.80	2.54 (0.58)	2.67 (0.62)	n.s.	2.79 (0.58)	2.59 (0.60)	n.s.
Control variables										
School total enrollment	515.22	167.58	179	957	507.56 (171.13)	522.87 (164.52)	n.s.	454.14 (120.01)	522.90 (171.38)	n.s.
School percent free/ reduced lunch	0.80	0.21	0.00	0.96	0.79 (0.22)	0.80 (0.20)	n.s.	0.87 (0.21)	0.79 (0.21)	n.s.
School percent White	0.11	0.20	0.00	0.90	0.11 (0.20)	0.12 (0.20)	n.s.	0.08 (0.12)	0.12 (0.20)	n.s.
School percent African American	0.71	0.30	0.00	0.99	0.72 (0.31)	0.69 (0.30)	n.s.	0.84 (0.21)	0.69 (0.31)	t(33) = 3.04

White, 2.6% were Asian American, 0.2% were Native American, and 6.7% were of another or mixed race/ethnicity. Approximately 51% of the student sample was female; and 52% of the student sample received a free or reduced-price lunch. Students were approximately evenly distributed across grade levels, with 24.9% in Kindergarten, 26.6% in 1st grade, 25.6% in 2nd grade, and 22.9% in 3rd grade.

2.3. Procedures

2.3.1. Recruitment and randomization

The school districts and project staff identified eligible teacher participants and conducted all recruitment sessions, either though attendance at new teacher training and orientation events or through individual or small group sessions. All teacher consenting processes were managed by project staff and participation was voluntary. All teachers provided written informed consent consistent with IRB procedures approved at the universities and school divisions. Participating teachers received a modest honorarium (e.g., gift cards) for their participation and completion of data collection activities. Identical recruitment and randomization procedures were used for each of the three cohorts of consenting teachers. Following baseline data collection, teachers were randomly assigned (within school district) to either receive training in GBG + MTP or serve as controls. Specifically, eligible teachers were recruited and randomly assigned to condition (blocking on school and district) using a random number generator. No additional supports or resources were provided as part of this RCT to others within the building for either condition, beyond a basic overview of the model for the principals in the participating schools.

2.3.2. Training in the GBG + MTP program

Professional development was provided to the intervention teachers over the course of two full days (weekends), during which participants received intensive training and practice using the PAX Good Behavior Game approach aligned with the MTP framework. Seventy-seven percent of the intervention teachers completed these two weekend trainings, with the remainder attending small group or one on one make-up trainings.

2.3.2.1. The PAX Good Behavior Game (PAX GBG). The GBG, originally developed by Barrish, Saunders, and Wolf (1969), allows teachers to utilize social learning principles within a team-based, game-like context to reduce aggressive/disruptive and off-task behavior and, consequently, facilitate academic instruction. In the current study we used the PAX GBG (Embry et al., 2003), which

represents efforts to improve the effectiveness of the original GBG and to make it ready for wide scale dissemination (Embry et al., 2003). As with the original GBG, the PAX GBG involves group-based rewards several times per day, using the Matching Law and Premack Principle (Knapp, 1976). Rotating student groups or "teams" are reinforced for their collective success in voluntary, group control contextually-defined "unwanted" behavior. The team-based nature of the game allows teachers to take advantage of positive peer pressure for improving academic and pro-social student behavior at the individual as well as at the classroom level. Prior to the implementation of the PAX GBG at the beginning of the school year, the teacher and students collaborate to define their vision of the ideal classroom. They identify the behaviors that they feel are necessary for creating a focused, productive, and peaceful classroom. During this collaboration, the teacher explains to the students that the positive behaviors they listed will be referred to as "PAX" behaviors and the negative behaviors will be referred to as "spleems." The use of the term, spleem, is designed to provide teachers an efficient, but emotionally neutral term, to refer to student misbehavior. After jointly defining what PAX and spleems are, teachers assign students to one of 3 to 4 teams. The teacher seeks to evenly match the teams in terms of student behavior to ensure all teams have an equal chance of winning the "game." The teams work cooperatively to maintain PAX behavior (which is Latin and intended to represent a state of peace, productivity, health, and happiness) in the classroom. Points are given to the team when a member displays a spleem. Teachers are trained to respond unemotionally to rule breaking and when marking points against a child's team. At the end of the game period, all teams with three or fewer spleems win the game. The students are rewarded for displaying self-control, emotional regulation and group regulation while not attending to or reinforcing the misbehavior of others. Additional elements of the PAX GBG to facilitate the development of self-control and generalization include verbal and visual cues that teachers and classmates can use to signal misbehaving, or off-task students, to stop spleeming.

2.3.2.2. MyTeachingPartner™. MTP introduces an evidence-based framework for thinking about teacher-student interactions that contribute to student behavior and achievement, called the CLASS (Hamre et al., 2012a); this framework concentrates on interactions that are emotionally supportive, well-organized, and cognitively enriching. The CLASS framework, including these three core domains of classroom interactions, is then used as a lens for viewing and providing feedback on videotapes of a teacher's practice in the classroom. Each teacher in the intervention condition was provided biweekly MTP coaching cycles throughout the training year, with initial contact in-person that then shifted to web-mediated training. Specifically, an MTP coaching cycle consists of the following steps:

1) The teacher tapes a 30-minute lesson including playing the GBG and sending this to his/her coach along with scoreboard records of all GBG "games" played since the last mailing.
2) The coach watches the video and edits out three key clips from that lesson/game for the teacher to watch. Video clips are accompanied by written prompts that help the teacher focus on one element of their practice, make specific behavioral observations about what is happening in the clip, and/or analyze how this is/is not working for their classroom.
3) The teacher watches these clips, reads the accompanying text, and reflects on and responds to prompt questions.
4) The coach reads teachers' responses and prepares for a conference.
5) The coach and teacher engage in a collaborative conference during which they identify successes/challenges, and brainstorm solutions and strategies related to teacher-student interactions and the GBG. During the conference, the coach and teacher select specific strategies for the teacher to try over the next few weeks that related to a specific teacher-student-interaction (e.g., teacher sensitivity) and GBG practices. These strategies will be the focus of the next coaching cycle.
6) The coach sends the teacher a summary and action plan outlining the topics covered in the conference and what the teacher is going to try in her classroom over the next two weeks and on the next videotaped lesson, as well as identifying supportive text and video resources (including models of effective interactions) that the teacher can access on the project website.

Over the course of the year for each teacher, these coaching cycles were intended to focus on all three CLASS domains as well as elements of the GBG that would help teachers optimize their implementation of the GBG by attending to their interactions with students. The coaching cycles were intended to be collaborative, supportive, constructive, and to help teachers develop CLASS and GBG knowledge, improve observation skills, develop analysis skills, and increase their sense of agency and efficacy in the classroom.

Teachers were asked to play approximately three GBG games each school day with increasing length and in increasingly varied settings over the course of the year. Teachers were also asked to incorporate new strategies related to both the CLASS domains and the GBG elements into their teaching practice in order to improve both their implementation of the GBG and their overall teaching practice.

2.3.3. Data collection

Pre-test survey of teacher experience and distress about teaching were collected at the outset of regular district training at the start of the school year and student and classroom assessments were collected in the fall semester, as close to the opening of school as feasible. Post-test assessments and ratings were obtained shortly before the end of the school year. School-level demographic data were obtained from the state.

2.4. Measures

2.4.1. Student achievement

Reading and math achievement were measured using two subtests for each area from the *Woodcock Johnson Tests of Achievement*

III (WJ III ACH; Woodcock, McGrew, & Mather, 2001). The WJ III ACH is designed to identify relative performance level on major academic areas. We used Reading Fluency and Reading Comprehension to assess reading, and Applied Problems and Mathematic Calculation to assess math achievement. We calculated two scores in accordance with guidelines in the manual, one for each subject and recorded the grade adjusted score at pre- and post-test for the subsample of 8–10 students selected from each classroom. These subtests have demonstrated high internal consistency reliability, and have been found to positively correlate with other measures of academic achievement (Woodcock et al., 2001). Given that students were in different grades, we applied grade adjusted scores to permit comparisons. As this test is intended to be able to measure change in achievement over time, by utilizing grade adjusted scores we could compare relative gain within the school year across classrooms of different grades (Dumont, Willis, & Walrath, 2016). Due to budgetary limitations, we were only able to collect these data for the students of the first two cohorts of teachers.

2.4.2. Classroom observations

Assessing School Settings: *Interactions of Students and Teachers* (ASSIST; Rusby, Crowley, Sprague, & Biglan, 2011). We used the ASSIST to assess students' classroom behavior, at one month into the school year and at posttest in the spring. The measure includes event-based tallies (e.g., counts of students' off-task behavior), and after the observation sessions, a set of global rating items, scored on a 5-point Likert-type scale, from 0 (*never*) to 4 (*almost continuously/often occurred*). The global rating items are averaged to create three scales: *student compliance* (seven items, e.g., "Students are focused and engaged"), *student socially disruptive behavior* (six items, e.g., "Social conversations occur between students and peers"), and *teacher and student meaningful participation* (nine items, e.g., "Students are provided opportunities to contribute to discussion"). In addition, observers provide a global estimate what proportion of students were off task. Previous research has shown high inter-rater agreement for each of the observation codes (ranging from 0.74 to 0.99; Rusby et al., 2011), and subscales have been validated in confirmatory factor analyses (see Debnam, Pas, Bottiani, Cash, & Bradshaw, 2015; Pas, Cash, O'Brennan, Debnam, & Bradshaw, 2015). In the present study, we utilized global rating scores reflecting the two student classroom behaviors of *student compliance* and *student socially disruptive behavior*, and the observers' global ratings of *students' off-task behavior*. These subscales have been validated in confirmatory factor analyses (e.g., Debnam et al., 2015; Pas et al., 2015). These global rating scores are less base-rate dependent than tally scores, avoid concerns around zero-inflation and skewed distributions, and are more readily comparable across classrooms and timepoints. Intraclass correlations (ICC) on the ASSIST scales ranged from 0.72 to 0.81, with an average of 0.75, indicating relatively little variability across the cycles. At pre- and post-test each classroom was observed on two separate days by trained observers, balancing afternoon and morning observations for a total of 4 to 6 cycles of the 15-min ASSIST observations per classroom. Scores reflect average rating across all available cycles at each time point.

Prior to data collection, all data collectors participated in a didactic training using a coding manual, videos and vignettes, followed by in-school practice observations with multiple expert coders. Thereafter, reliability was assessed by the degree to which trainees' codes matched the expert coders' codes for the classroom (i.e., inter-observer agreement). Observers were expected to achieve an 80% inter-observer agreement across three practice observations to be considered reliable. After achieving 80% agreement, observers coded independently in each classroom, with an average reliability ranging from 85 to 88% across data collection waves. Prior reliability analyses of the ASSIST further indicate high reliability among observers, as a very low proportion of the variance (<1%) in the classroom codes was attributable to the independent raters (Abry, Cash, & Bradshaw, 2014).

2.4.3. Teacher distress

A composite score reflecting teachers' *distress related to teaching* was computed as a mean of nine items drawn from two existing scales. Four items measured teachers' feelings of emotional exhaustion and disillusionment about teaching (e.g., "I feel emotionally drained from my work;" "I feel burned out from my work;" taken from the Maslach Burnout Inventory [Maslach, Jackson, & Leiter, 1996]). Five items were taken form the National Institute for Occupational Safety and Health survey of work-related stress (National Institute for Occupational Safety and Health, 1999) and reflect teachers' work stress effects (e.g., "I regularly experience physical symptoms [e.g., headache, stomach upset, muscle tension] associated with stress;" "I am unable to cope with the stress of my job on a daily basis"). We combined these items as each reflect distress due to work requirements. Also, given the sample was only teachers at or near the beginning of their career we wanted to capture more than burnout as typically measured (which reflects primarily disengagement due to experience over time). To ensure the scale was valid, we conducted scaling analyses which showed that all items loaded and the scale had good internal consistency (α = 0.87–0.90 across cohorts). All items had responses scaled from 1 to 4 (*strongly disagree* to *strongly agree*), with scores based on the average of the non-missing items, such that higher scores reflect greater distress. Instructions referred to reflecting on the whole of the prior teaching experience.

2.4.4. Control variables

Teachers self-reported on their gender, race/ethnicity, education, and years of experience. In addition, school total enrollment, student racial/ethnic distribution, and free or reduced lunch rate, as well as classroom grade level, were recorded from annual school record searches.

2.5. Analytic Plan

Although there was some clustering of teachers within schools, the average number of teachers per school was very small (2.32), with only one teacher per school in about 44% of schools. With intraclass correlations (ICCs; representing the proportion of variance in outcomes that is between schools) ranging from 0.01 to 0.26 across outcomes, the "design effect" (1 + (average cluster size −1) *

ICC) falls well below 2 (ranging from 1.01 to 1.34) across all outcomes, suggesting minimal bias to standard errors due to clustering of teachers within schools (Muthén & Satorra, 1995). Thus, training effects on class mean student behavior and achievement, including potential moderation of intervention effects, were assessed through a series of single-level, multiple linear regression analyses computed in MPlus version 8 (Muthén & Muthén, 2017). Full information maximum likelihood (FIML) was used to incorporate cases with missing data. Student missing cases were not handled with FIML prior to aggregating; we simply took the mean score across students with available data for each teacher (ranging from 2 to 14 students per teacher). Of teachers for whom we had student data, 97% had student data from at least 5 students, and 84% had student data from at least 10 students. FIML was used in all models, including the models predicting our achievement outcomes (for which we have data from 95 teachers from Cohorts 1 and 2).

Each regression model tested one of the five targeted class mean student outcomes at post-test (i.e., student socially disruptive behavior, student off-task behavior, student compliance, reading achievement, math achievement), predicted by intervention condition and its interaction with baseline teacher distress and baseline classroom student socially disruptive behavior, including the three-way interaction among these variables and each of the main effects and two-way interactions that make up the three-way interactions. All models controlled for corresponding pre-test (baseline) scores, as well as teacher and student demographics.

3. Results

3.1. Descriptive analyses

Descriptive statistics (means, distributions) for control, predictor, moderator, and outcome variables are listed in Tables 1 and 2, with bivariate correlations provided in Table 3. In addition, Tables 1 and 2 provide mean comparisons by condition at baseline, as well as baseline comparisons of teachers who dropped out after randomization to those who remained in the study at post-test.

3.2. Implementation fidelity

Fidelity was assessed using teacher self-reports and ratings by independent observers of videotapes of teachers implementing the GBG + MTP procedures**.** Teachers recorded the number of times they implemented/played the GBG and the length of the games played. On average, participating teachers played the GBG 9–16 ($SD = 2.81$) times per week within a range of 2–15 games per week and for 76.26 min ($SD = 48.32$) per week within a range of 15–316 min. Fidelity of GBG implementation was primarily measured using a rubric developed and used in prior GBG randomized trials (Domitrovich et al., 2015). Videos of the teachers during planned GBG implementation were rated 0–4 for quality of implementation as intended, by independent raters (interrater reliability was 0.79 for Fall and 0.89 for Spring). Scores are the average across 6 observations at each data point. For Fall observations the mean quality score on a scale of 0–4 was 3.27 ($SD = 0.39$) and for Spring was 3.25 ($SD = 0.48$).

To track MTP component implementation and fidelity, we used an online system that recorded the number of completed MTP coaching cycles. On average, teachers participated in 8.28 coaching cycles ($SD = 2.39$), ranging from 1 to 11 with a median of 9. These average cycle numbers met and exceeded the target of 8, which is the lowest dose of MTP cycles previously shown in a RCT to impact teacher practice and student outcomes (Allen et al., 2011; Allen et al., 2015). In addition, coaches reported that on average conferences with teachers lasted 32.83 min ($SD = 9.21$), which is well-aligned with the intended 30-minute target. On a scale from 1 (*Strongly Disagree*) to 4 (*Strongly Agree*), coaches also indicated that across cycles they had productive conferences with teachers ($M = 3.73$, $SD = 0.51$) and that teachers understood how to use the CLASS framework to discuss their classroom practices ($M = 3.61$, $SD = 0.55$). On the same scale, teachers rated their experience of MTP coaching; they agreed that the online prompts provided by the coach were relevant to their practice ($M = 3.65$, $SD = 0.54$) and that answering the prompts and watching the video was worth the time it took ($M = 3.60$, $SD = 0.58$).

3.3. Effects analyses

As reported in Tables 4 and 5, across all five outcomes, there was a pattern of statistically significant effects (or in one case, a marginally significant effect) for the three-way interaction of intervention condition with baseline teacher distress and baseline classroom socially disruptive behavior. Specifically, the three-way interaction of intervention condition, baseline teacher distress, and baseline student socially disruptive behavior significantly predicted relative pre- to post-test growth in student socially disruptive behavior ($B = -0.57$, $p < .001$), student off-task behavior ($B = -0.38$, $p = .03$), student compliance ($B = 0.34$, $p = .02$), math achievement ($B = 0.09$, $p = .05$), and reading achievement ($B = 0.09$, $p = .05$). In addition, two of the models had a significant two-way interaction of teacher distress by intervention condition. Specifically, baseline teacher distress significantly moderated the effect of intervention condition on student socially disruptive behavior ($B = -0.28$, $p = .03$) and student off-task behavior ($B = -0.29$, $p = .04$), though these two-way interactions were only significant when the three-way interaction was also included in the model. There were no main effects of intervention condition in any of the models.

To facilitate interpretation of the significant moderation effects noted above, we computed (and plotted, in Figs. 2–6) the simple slopes reflecting relations between intervention condition and each outcome at specific conditional values of our moderators: baseline classroom socially disruptive behavior and baseline teacher distress (or just baseline teacher distress, for the significant two-way interactions) (Preacher, Curran, & Bauer, 2006). Specifically, we calculated the median score of classrooms at pre-test in the top third ($n = 62$) of the overall sample on each respective moderator (referred to as “high”) and the median score of all other classrooms (lower two-thirds; $n = 120$; 6 participants were missing values on the moderators) on each respective moderator (referred to as

Table 3
Bivariate correlations.

	Baseline						Post-test				
	Teacher distress	SSDB	Off-task	Compliance	WJ reading	WJ math	SSDB	Off-task	Compliance	WJ reading	WJ math
Teacher demographics											
Female	0.04	−0.07	−0.07	−0.02	−0.07	−0.13	−0.03	0.04	−0.04	−0.09	−0.11
Non-white	−0.07	−0.03	−0.07	0.01	n/a	n/a	−0.01	0.01	−0.06	n/a	n/a
Year of teaching	0.07	**−0.28****	**−0.16***	**0.20****	−0.16	−0.13	−0.15	−0.06	0.140	−0.15	−0.09
Master's degree	0.08	0.01	−0.11	−0.04	0.03	0.08	0.06	0.13	−0.08	0.00	0.06
Grade level	0.16*	0.03	**0.16***	−0.05	**0.82****	**0.84****	0.01	**0.15***	<0.01	**0.81****	**0.83****
School-level demographics											
Total enrollment	−0.01	−0.10	0.06	0.02	0.06	0.01	−0.06	−0.03	0.02	0.08	0.03
Percent free/reduced lunch	0.11	0.14	0.14	**−0.25****	**−0.36****	**−0.30****	**0.31****	**0.24****	**−0.40****	**−0.39****	**−0.34****
Percent White	−0.06	−0.09	−0.13	**0.22****	**0.34****	**0.30****	**−0.21****	−0.09	**0.30****	**0.36****	**0.33****
Percent African American	0.10	**0.26****	**0.18***	**−0.34****	**−0.30****	**−0.26***	**0.28****	**0.13**	**−0.32****	**−0.32****	**−0.29****
Moderators											
Teacher distress	–	0.12	0.18*	−0.23**	<0.01	0.03	0.08	0.16*	−0.14	−0.02	0.01
Student socially disruptive behavior (SSDB)	0.12	–	0.57**	−0.71**	−0.01	0.02	0.53**	0.33**	−0.50**	−0.04	−0.03
Baseline scores on outcome variables											
Student off-task behavior	0.18*	0.57**	–	−0.66**	0.22*	0.22*	0.48**	0.43**	−0.51**	0.20	0.18
Student compliance	−0.23**	−0.71**	–0.66**	–	−0.08	−0.12	−0.61**	−0.49**	0.66**	−0.05	−0.06
WJ reading composite	0.00	−0.01	0.22*	−0.08	–	0.96**	−0.07	0.09	0.07	0.98**	0.96**
WJ math composite	0.03	0.02	0.22*	−0.12	0.96**	–	−0.02	0.12	0.00	0.94**	0.98**
Post-test scores on outcome variables											
Student socially disruptive behavior (SSDB)	0.08	0.53**	0.48**	−0.61**	−0.07	−0.02	–	0.66**	−0.81**	−0.12	−0.09
Student off-task behavior	0.16*	0.33**	0.43**	−0.49**	0.09	0.12	0.66**	–	−0.68**	0.04	0.06
Student compliance	−0.14	−0.50**	−0.51**	0.66**	0.07	0.00	−0.81**	−0.68**	–	0.11	0.07
WJ Reading Composite	−0.02	−0.04	0.20	−0.05	0.98**	0.94**	−0.12	0.04	0.11	–	0.96**
WJ Math Composite	0.01	−0.03	0.18	−0.06	0.96**	0.98**	−0.09	0.06	0.07	0.96**	–

Note. Bolded values are significant, $*p < .05$, $**p < .01$, $***p < .001$

Table 4
Effects of GBG + MTP on student behavior.

	Student Socially Disruptive Behavior			Student Off-Task Behavior Count			Student Compliance		
	Estimate (SE)	ΔR2	Total R^2 (SE)	Estimate (SE)	ΔR2	Total R^2 (SE)	Estimate (SE)	ΔR2	Total R^2 (SE)
Intercept covariates	−0.10 (0.53)			−0.67 (0.56)			0.39 (0.47)		
DV at pre-test	–			0.27** (0.09)			0.40*** (0.09)		
Gender	0.09 (0.25)			0.36 (0.27)			−0.20 (0.22)		
Non-white	−0.30 (0.45)			0.07 (0.49)			−0.21 (0.40)		
Years of experience	−0.07 (0.07)			<0.01 (0.07)			0.05 (0.06)		
Level of education	0.17 (0.15)			0.46** (0.16)			−0.14 (0.13)		
District_BaltoCo	−0.32 (0.20)			−0.38+ (0.22)			0.13 (0.18)		
District_PGCo	0.47 (0.36)			−0.07 (0.38)			−0.48 (0.32)		
Total enrollment	0.02 (0.08)			−0.04 (0.09)			0.03 (0.07)		
Free/reduced lunch (log transformed)	−2.33 (1.62)			−0.71 (1.73)			2.83* (1.42)		
% African American	0.26+ (0.16)			−0.02 (0.17)			−0.18 (0.14)		
% White	1.97 (2.28)			−0.23 (2.44)			−0.61 (2.02)		
First grade	0.02 (0.18)			<0.01 (0.19)			−0.09 (0.15)		
Second grade	0.16 (0.19)			0.23 (0.20)			−0.22 (0.17)		
Third grade	0.16 (0.18)			0.37+ (0.20)			−0.14 (0.16)		
Time between pre- and post-test	0.04 (0.07)			0.01 (0.08)			0.11 (0.07)		
Treatment group (group)	0.03 (0.12)			−0.09 (0.13)			−0.10 (0.11)		
Statistics for Step I		*0.387*	*0.387* (*0.10*)		*0.282*	*0.282* (*0.06*)		*0.523*	*0.523* (*0.08*)
Moderators									
Student socially disruptive behavior (SSDB) at baseline	0.57*** (0.10)			0.32** (0.12)			−0.24* (0.11)		
Teacher distress at baseline	0.09 (0.09)			0.23* (0.10)			−0.06 (0.08)		
Group X SSDB	−0.17 (0.13)			−0.25+ (0.14)			0.13 (0.12)		
Group X teacher distress	−0.28* (0.13)			−0.29* (0.14)			0.18 (0.11)		
SSDB X teacher distress	0.23* (0.10)			0.32** (0.11)			−0.18* (0.09)		
Statistics for Step II		*0.006*	*0.393* (*0.09*)		*0.046*	*0.328* (*0.07*)		*<0.001*	*0.521* (*0.07*)
Group X SSDB X teacher distress	−0.57*** (0.16)			−0.38* (0.17)			0.34* (0.14)		
Statistics for Step III (3-way interaction)		*0.071*	*0.464* (*0.10*)		*0.011*	*0.339* (*0.06*)		*0.03*	*0.551* (*0.08*)

Note. + $p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

"low") and used the conditional values in the computation of simple slopes; we then explored whether classrooms in the intervention condition differed from those in the control condition within each of these "subgroups". For example, for the three-way interactions, we estimated the effects for classrooms with high student disruption and high teacher distress at baseline, those with high disruption and low distress at baseline, etc. Given the challenges inherent in interpreting effect sizes of 3-way interactions from squared semi-partial correlation values, which are reported in Tables 4 and 5 (Aiken & West, 1991; McClelland & Judd, 1993), we additionally computed Cohen's *d* effect sizes for the model-implied mean difference between the intervention and control group *within* each of these subgroups. Specifically, we used the difference between the intervention and control of each subgroup in their respective estimated values on each outcome at our selected conditional values of each moderator, divided by the pooled standard deviation of the outcome within the respective subgroup (reported below). The three-way interaction effects are graphed in Figs. 2–6, which are presented such that adjusted post-test means (controlling for pre-test) on each outcome are plotted as a function of condition (on x-axis) by levels of baseline teacher distress (dotted line for high distress, solid line for low distress) for highly disruptive classrooms in the top graph, and for low disruptive classrooms in the bottom graph. The outcome of interest is listed on the left side of the Y-axis.

Review of these illustrative simple slopes and plots indicate that the intervention effects were primarily limited to classrooms that were high in disruptive behavior at baseline and had teachers who expressed high distress about teaching at baseline. For the two-way interactions, comparison of the simple slopes (representing the relation between intervention condition and student outcomes) in the high versus low teacher distress subgroups suggest greater benefits of training (intervention) for teachers that were high in distress at baseline than for low distress teachers. For students' off-task behavior, a significant simple slope ($B = -0.38, p = .047$) for high distress teachers suggests a significant difference between the intervention and control condition among teachers who were high in distress at baseline. For students' disruptive behavior, the simple slopes were *not* statistically significantly for either of the subgroups (low or high distress teachers); thus, while intervention condition is more positively associated with students' disruptive behavior among high distress teachers than among low distress teachers, the difference between intervention and control teachers is not statistically significant in either of these subgroups. It is important to note, however, that the two-way interaction of intervention condition by teacher distress was only statistically significant conditional upon the inclusion of the three-way interaction in the

Table 5
Effects of GBG + MTP on student achievement.

	Reading composite			Math composite		
	Estimate (SE)	ΔR2	Total R^2 (SE)	Estimate (SE)	ΔR2	Total R^2 (SE)
Intercept covariates	−0.03 (0.18)			−0.28 (0.19)		
DV at pre-test	0.95*** (0.05)			0.93*** (0.06)		
Gender	<0.01 (0.07)			0.10 (0.08)		
Non-white	−0.97** (0.36)			1.27** (0.38)		
Years of experience	−0.01 (0.02)			0.07** (0.02)		
Level of education	−0.08 (0.05)			−0.09+ (0.05)		
District_BaltoCo	−0.06 (0.05)			0.08 (0.05)		
District_PGCo	–			–		
Total enrollment	0.01 (0.02)			−0.03 (0.03)		
Free/reduced lunch (log transformed)	0.77* (0.39)			0.34 (0.41)		
% African American	0.01 (0.05)			−0.07 (0.05)		
% White	0.20 (0.66)			−0.45 (0.67)		
First grade	0.01 (0.07)			<0.01 (0.08)		
Second grade	−0.08 (0.09)			0.05 (0.11)		
Third grade	0.04 (0.13)			0.12 (0.17)		
Time between pre- and post-test	0.05* (0.02)			0.03 (0.02)		
Treatment group (group)	0.03 (0.03)			−0.01 (0.04)		
Statistics for Step I		*0.980*	*0.980 (<0.01)*		*0.980*	*0.980 (<0.01)*
Moderators						
Student socially disruptive behavior (SSDB) at baseline	0.02 (0.03)			0.03 (0.03)		
Teacher distress at baseline	−0.08** (0.03)			−0.06+ (0.03)		
Group X SSDB	−0.01 (0.04)			−0.03 (0.04)		
Group X teacher distress	0.06 (0.04)			0.04 (0.04)		
SSDB X teacher distress	−0.09** (0.03)			−0.06* (0.03)		
Statistics for Step II		*0.004*	*0.984 (<0.01)*		*0.003*	*0.983 (<0.01)*
Group X SSDB X teacher distress	0.09+ (0.04)			0.09* (0.05)		
Statistics for Step III (3-way interaction)		*<0.001*	*0.984 (<0.01)*		*0.001*	*0.984 (<0.01)*

Note. + $p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

model; thus, the two-way interactions need to be interpreted in the context of the three-way interactions.

For the three-way interactions, significant simple slopes for the high distress, highly disruptive classrooms suggest that training (intervention) provided significant benefits for this "highest risk" subgroup across four of five tested outcomes. Specifically, highly disruptive classrooms with highly distressed teachers (at pre-test) in the intervention condition showed significantly more favorable adjusted post-test means (after controlling for corresponding baseline scores) than did such classrooms in the control condition, with regard to student socially disruptive behavior (simple slope $B = -0.89$, $p < .001$; Cohen's d for the model-implied mean difference between intervention and control among these classrooms = 0.70; see Fig. 2), student off-task behavior ($B = -0.92$, $p < .01$; $d = 0.54$; Fig. 3), student compliance ($B = 0.47$, $p < .05$; $d = 0.45$; Fig. 4), and reading achievement ($B = 0.15$, $p < .05$; $d = 0.18$; Fig. 5). In contrast, non-significant simple slopes for each of the other subgroups of classrooms (highly disruptive classrooms with low distress teachers, low disruptive classrooms with high distress teachers, and low disruptive classrooms with low distress teachers) suggested no difference in these outcomes between the intervention and control conditions within any of these other subgroups.

Although the three-way interaction of intervention condition with baseline teacher distress and student disruption was statistically significant in predicting math achievement, none of the simple slopes at our conditional values of teacher distress and disruptive behavior were statistically significant. Considered alongside the plots of our simple slopes (Fig. 6), the significant three-way interaction suggests that within highly disruptive classrooms, the relation between intervention condition and outcome was relatively more positive among high distress teachers, and relatively more negative among low distress teachers (i.e., these two slopes differed from each other). But the non-significant simple slopes indicate that this relation did not significantly differ from zero for either group; or, put differently, adjusted post-test mean math achievement did not significantly differ between intervention and control within any of the sub-groups examined ($d = 0.11$ for the difference in math achievement between intervention and control within the highest risk subgroup).

Across all outcomes, the adjusted post-test mean plots suggested that within highly disruptive control classrooms, students with high distress teachers had worse adjusted post-test scores than those with low distress teachers on each behavioral and academic outcome. The differential patterns of growth by condition suggest that for off-task behavior as well as reading and math achievement, this difference between high and low distress teachers in highly disruptive classrooms was eliminated in the intervention group; for student socially disruptive behavior and compliance, high distress teachers in highly disruptive intervention classrooms actually outperformed their low distress counterparts. In other words, it appears that the intervention served a protective effect for higher risk classrooms, helping highly distressed teachers facing highly disruptive classrooms to improve (or prevent worsening of) behavior and achievement more than would occur otherwise, reaching similar adjusted post-test levels to those seen in the lower risk subgroups. While the squared semi-partial correlations (as shown in Tables 4 and 5) suggest a small percent of variance uniquely explained by the three-way interaction in each of our models (ranging from 0.001 to 0.07 across outcomes, values that are common for interaction

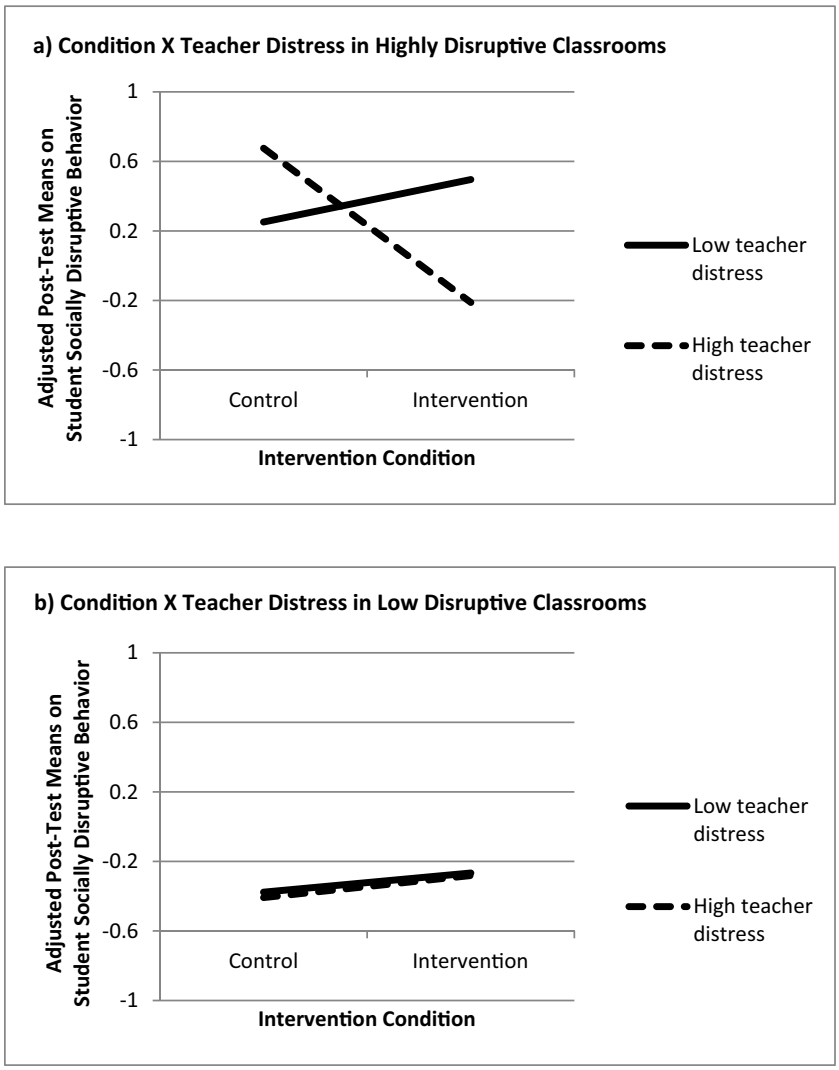


Fig. 2. Adjusted post-test means (controlling for corresponding pre-test scores) for student socially disruptive behavior by intervention condition and teacher distress level, within classrooms that were high (a) vs low (b) on student socially disruptive behavior at baseline.

effects; Aiken & West, 1991), Cohen's *d*'s computed on the model-implied mean difference between intervention and control group within the highest-risk subgroup suggest effects of the intervention condition within this subgroup may be small for achievement outcomes, but medium to large for behavioral outcomes.

4. Discussion

This study focused on the effects on student outcomes of training early career teachers in two evidence-based approaches for classroom organization and behavior management skills. We also examined how student effects might be conditioned on teachers' distress as they start teaching and how challenging the classroom was they had to manage. The results of these moderation analyses may help inform training content and targeting of professional development activities and resources.

Prior randomized control trials of teacher training, including preventive efforts, have included teachers irrespective of years of teacher experience. However, our findings highlight the potential significance of “preventive training” in improving novice teachers' behavior management, particularly in high-risk circumstances. In addition to bringing recognition that teacher training needs are likely to vary as a function of experience level, this study illustrates the importance of considering experience in evaluating uptake and impact of training programs on student outcomes. A related consideration is how training programs might be modified based on the experience level of teachers; more experienced teachers need perhaps different programs. The training provided in this study was in fact focused on the concerns, needs, and skills of new teachers, particularly the often noted need for support related to classroom management (Browers & Tomic, 2000).

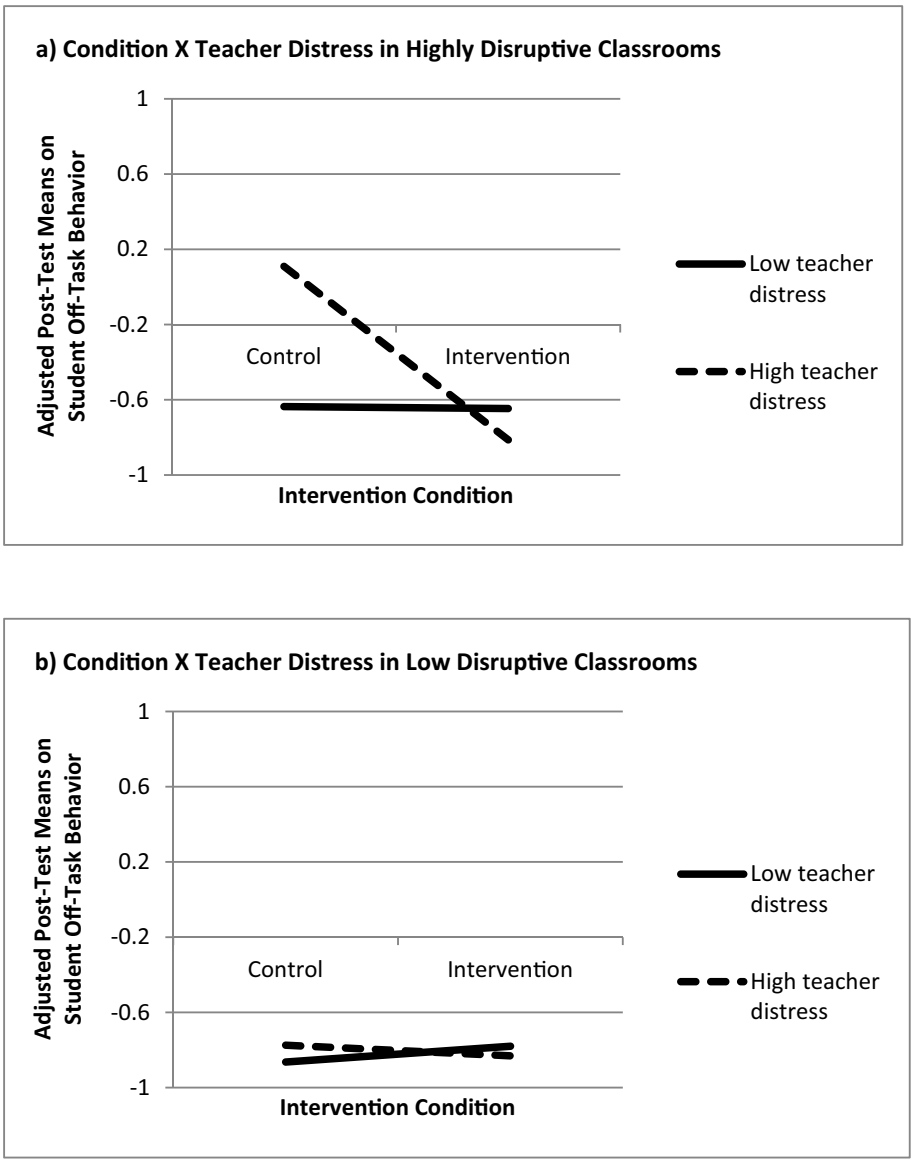


Fig. 3. Adjusted post-test means (controlling for corresponding pre-test scores) for student off-task behavior by intervention condition and teacher distress level, within classrooms that were high (a) vs low (b) on student socially disruptive behavior at baseline.

However, this seemingly valuable consideration may need tempering as the results here do not suggest heightened impact after one year of training and coaching compared to RCT evaluations that have engaged teachers irrespective of experience. There were no main effects for exposure to the combined training programs. However, this is not uncommon in prior studies of both GBG and MTP, as findings of moderated effects and concentration of impact tend to be greatest among the highest risk students (Kellam et al., 1994; Mikami et al., 2011). Moreover, while there was a couple of findings suggesting benefits may extend to initially distressed teachers irrespective of initial classroom disruption level, those effects only occurred with consideration of the impact of the three-way interactions. The present findings suggest the effects may be greatest not only for the highest risk students but also for classrooms with more overall behavior problems. Moreover, the effects seem to eventuate for when these classrooms are assigned to the higher risk"teachers who begin the school year high in distress (and thus may be especially in need of support for managing the classroom). The medium to large effects on behavior – and small effects on reading achievement – within this subgroup of high distress teachers with highly disruptive classrooms and consistency of the pattern increases confidence the findings are substantial.

Given the variations in features of this RCT from most prior teacher training evaluations, direct comparison to findings of prior evaluations is limited. In addition to inclusion of teachers irrespective of experience level, most prior trials have focused on variation in effects within classrooms across teachers, showing effects tend to concentrate among the students with elevated misbehavior. Our focus here is on differences between classroom levels of disruption, on-task, compliance, and achievement. This different design focus

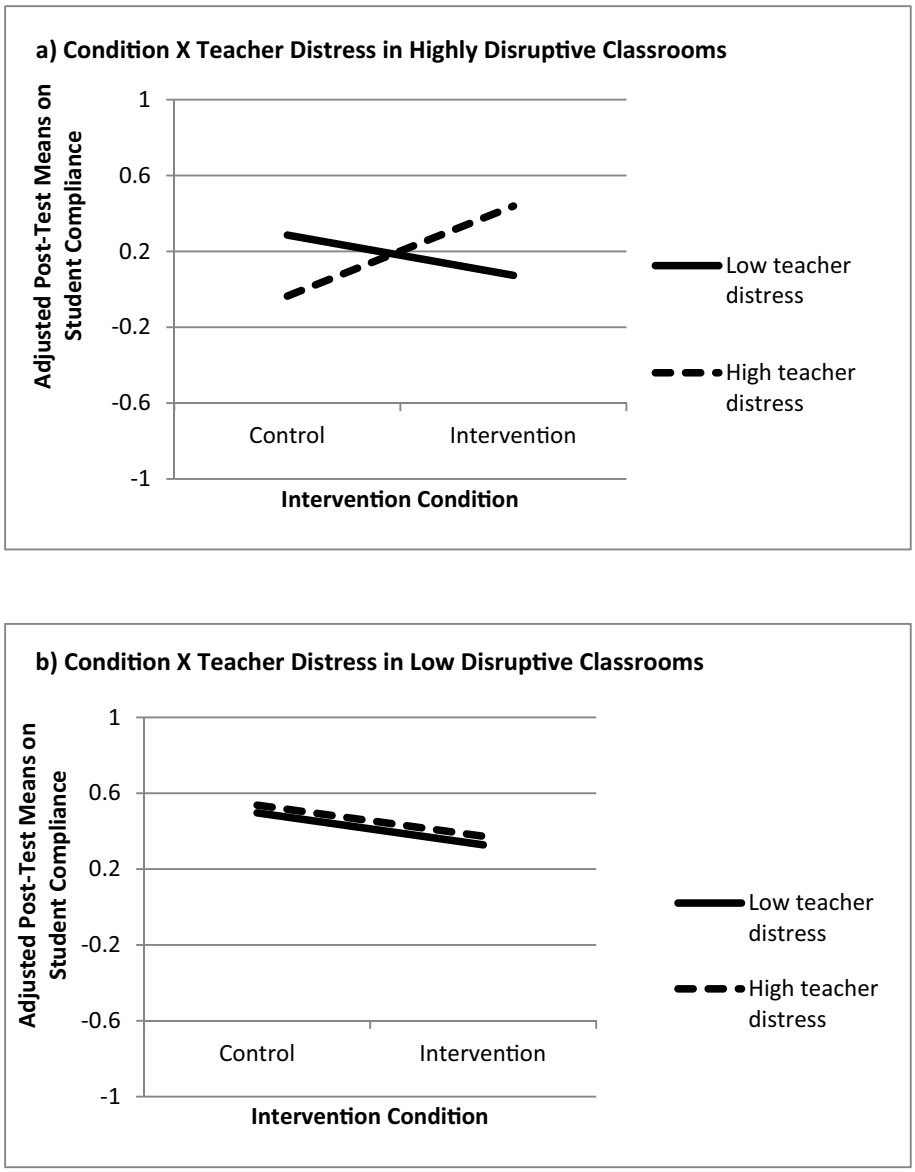
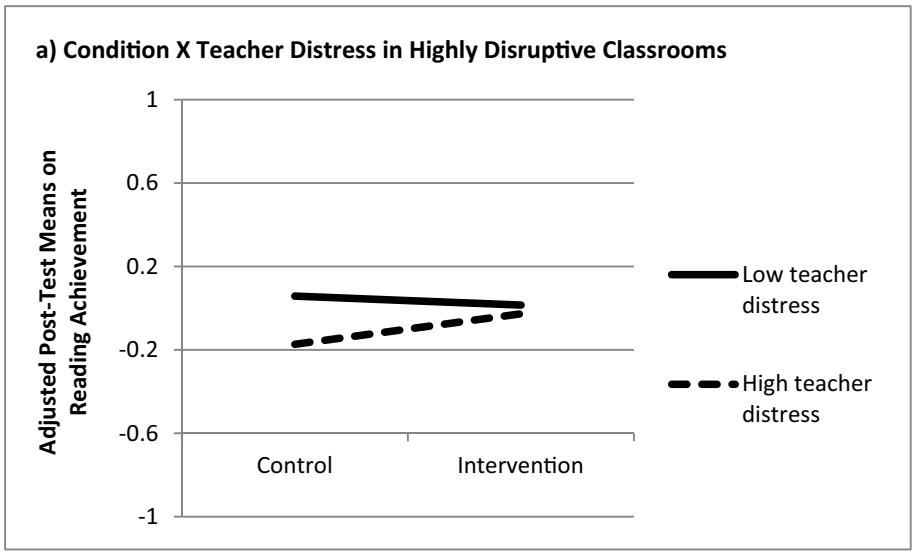


Fig. 4. Adjusted post-test means (controlling for corresponding pre-test scores) for student compliance by intervention condition and teacher distress level, within classrooms that were
high (a) vs low (b) on student socially disruptive behavior at baseline.

(at least in comparison to studies to date) permits a unique understanding, highlighted by the variation in effects related to teacher distress and classroom level of management challenge.

This study applied consideration of two potential conditional influences on training effects that have been given substantial mention in descriptive and theoretical formulations about teacher training effects, but have not been tested as moderators of training in prior trials (Evertson & Weinstein, 2006; Jennings & Greenberg, 2009). The first represents teacher person effects on training and how individual differences in comfort with and felt preparedness for teaching affects training impact. Those starting teaching bring different levels of confidence, skills, and motivation which likely affect how well they manage classrooms, but also how they might respond to training. We included this consideration, through assessment of teacher distress about the tasks and demands of teaching early in careers. The second represents setting effects or variation in the classroom teachers are facing, through variation in student disruptiveness from the outset of the year. As teachers face variation in how challenging managing his or her classroom at or near the outset of their career, it seems valuable to test how this might affect training impact. Yet, these two influences have not been heretofore considered as moderators of training impact, separately or in combination. The results suggest that both have importance in determining immediate benefit of training for student behavior and achievement. This pattern of findings raises important questions for further study, such as whether training in classroom management might be most efficiently focused on those with the



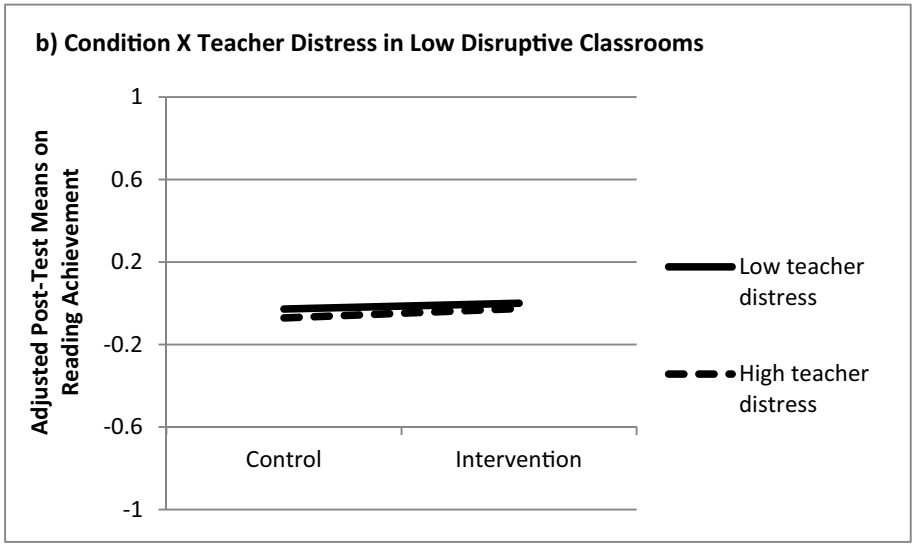


Fig. 5. Adjusted post-test means (controlling for corresponding pre-test scores) for student composite reading scores on the Woodcock Johnson Tests of Achievement III (WJ III ACH; Woodcock et al., 2001) by intervention condition and teacher distress level, within classrooms that were high (a) vs low (b) on student socially disruptive behavior at baseline.

most challenging classrooms and/or showing higher distress as they begin teaching.

The results of this baseline to end of year evaluation suggest, with substantial consistency, that the beneficial effects are limited to the interaction of highly distressed teachers and highly disruptive classrooms as initial conditions. It appears, for new teachers, the training provided here, helps those facing the more challenging classrooms and who also are engaging in teaching with greater distress about ability to manage classrooms and accomplish educational goals for their students. The effects are most clear and consistent (with the largest effect sizes) for student behavior, which is expected as behavior management is the central focus of training across the intervention components. Yet there is also evidence of promising (though small) impact on academic achievement in these same high-risk classrooms. Across three indicators of student behavior (disruptive behavior, off-task, compliance) and reading achievement this pattern emerges, lending confidence that the dual moderation differentiation of effects is robust.

Transfer of classroom management and organization training to student achievement and not just affecting student behavior, while often sought and theorized, has not been demonstrated often (Evertson & Weinstein, 2006; Luiselli, Putnam, Handler, & Feinberg, 2005). The results provide empirical support for training leading to improved student performance, eliminating the gap that emerges otherwise. This linkage suggests one target for reducing disparities in achievement and learning gains might be aiding

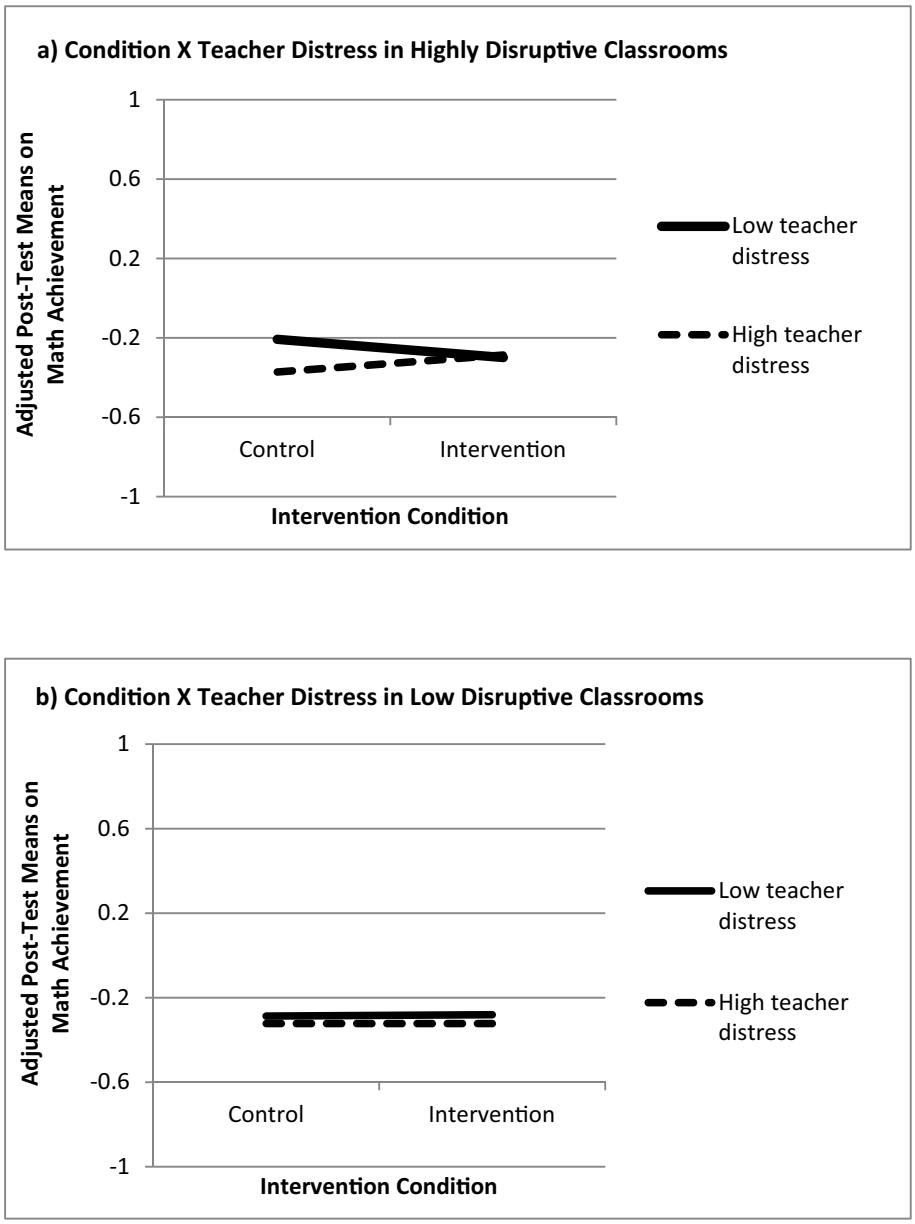


Fig. 6. Adjusted post-test means (controlling for corresponding pre-test scores) for student composite math scores on the Woodcock Johnson Tests of Achievement III (WJ III ACH; Woodcock et al., 2001) by intervention condition and teacher distress level, within classrooms that were high (a) vs low (b) on student socially disruptive behavior at baseline.

novice teachers with classroom management, particularly those in classrooms with an initially high level of behavior problems. While suggesting that classroom management training can lead to academic gains for students in disruptive classrooms with distressed teachers, these results cannot determine if the results are due to more opportunity for learning with improved classroom conditions or more learning due to better behavior among students. Further study to trace the path of such linkage may enhance further training efforts and improve this translation.

An additional innovation of this study is to combine two empirically tested teacher training programs for classroom management with augmentation of critical principles of classroom organization culled from teacher training literature. There are relatively few empirically tested teacher training programs particularly focused on the critical issues of classroom management. Based in a developmental (early intervention) and ecological (person, situation influences) conception of teacher training, the results can inform not only the value of the specific training, but also to suggest a more complex conceptualization of teacher training to facilitate student behavior and achievement. The program combines the approach of MTP that emphasizes quality of relationship with each student and overall classroom climate promotion for instructional opportunity with the group reinforcing behavior training of GBG.

The effects are similar to other trials for the GBG and MTP in finding moderated benefits that concentrate in the most at-risk portion of the students (Ialongo et al., 1999; Mikami et al., 2011). However, like most trials of teacher-based prevention, these prior efforts were designed to evaluate effects on individual students not on the classroom overall; the moderated effects typically examined have occurred at the level of the individual, rather than at the classroom. In the current study, the moderation of effects was not a test of differential impact for highest risk *individuals within* included classrooms, but rather of differences between higher risk *classrooms*. Thus, this study adds to the growing body of literature suggesting that the level of behavior problems may moderate the effects of training. This, in turn, may inform future deployment of schools' typically limited training and coaching resources, opting to concentrate them on those higher risk contexts, which may return the most benefit.

The programs were integrated because surveys and reviews of teacher training formulations and empirical studies suggest each contributes to more effective classroom management. While showing some promising results, this was not a direct comparison of the GBG and MTP individual components or meant to be a determination of whether the combination increases impact. Given the difference in design of this trial and most others (e.g., selection of early teachers, focus on classroom level student outcomes), it is difficult to compare relative effect sizes with much certainty of what these reflect. It is notable that while there is a prevalence of attention in the descriptive literature to classroom organization principles, this is, as far as we could find, the first randomized empirical test of training that incorporates specific emphasis on teachers at this stage. An additional consideration in judging the value of this particular training program is that year-long coaching was provided to teachers. This feature is central to MTP and a core element of many prior trials that provided evidence of GBG benefits (e.g., Ialongo et al., 1999). The impact that comes from regular and systematic support and feedback may be as important and cannot be differentiated in this case from the training content and techniques taught.

5. Limitations and future directions

In addition to limitations in differentiating training components and how training links behavior improvement to achievement, this study is limited by some methodological features. The sample size while large for a study of teacher training is not as large as would be preferred for testing moderation, affecting power to detect all effects. Also, while in line with typical training efforts, the participation and retention rates affected sample size and sample completeness. The districts from which teachers were engaged tended to serve more impoverished and economically disadvantaged students compared to what is typical in this country. This may affect confidence of generalizability to different populations. However, this study focused on a population for which there is great need for evidence-based programming and understanding about how to support teachers and promote student behavior and achievement.

Taken together, these results provide promising support for the potential impact of the integrated GBG and MTP program for novice teachers. This training program could be considered selective in that teachers were included based on little or no prior experience. However, it could also be considered as a universal program as it was organized for teachers at this stage irrespective of any other characteristics. Further, the results bring into focus consideration of further selection of teachers; perhaps initial distress might be a basis for selection or training might be applied based on empirical measurement of the classroom initial disruptive level. Empirical testing of whether further selectivity or this universal inclusion of new teachers is preferable seems like a valuable next step in improving precision of inclusion and potentially increasing impact of resource allocation (e.g., focusing on high risk teachers or situations among those at the outset of career vs. applying universally).

Although showing consistent findings for the buffering effects in the higher risk settings, there are some inconsistencies in the findings, and the overall effect sizes were relatively modest in many cases. At the same time, the study has incorporated multiple features that are often discussed as important in affecting training but have not been given much empirical attention before. Perhaps, as was the impetus for the trial, more complex consideration of teachers' developmental stage, teacher individual variation, and setting demands will be considered in future trials, both in terms of potential effect modifiers and as factors to target in developing and deploying interventions. Similarly, these are post-test results, so it would be valuable to undertake longer term analyses to see if these patterns are sustained.

In conclusion, the findings highlight the potential value of more attention to teacher distress and of classroom conditions as moderators of training effects, at least among novice teachers. It seems that consideration of both teacher predisposition and of the variation in how difficult the classrooms they first are assigned should inform the tailoring of training and coaching supports, as well as help one gauge the expected effects of professional development efforts. Moreover, consideration of these aspects of the ecology of teaching may enable more effective consultation and other forms of aid for teachers, improving the impact of these evidence-based approaches.

Acknowledgments

This work is supported by a grant from the Institute for Educational Sciences (R305A130107). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Education.

References

Abry, T., Cash, A. H., & Bradshaw, C. P. (2014). *Using generalizability theory to examine sources of variance in observed behaviors within high school classrooms.* Washington, DC: Unpublished report presented at the Society for Research on Educational Effectiveness.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* New York: Sage Publications, Inc.

Allen, J. P., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the MyTeachingPartner-secondary intervention. *Journal of Research on Educational Effectiveness, 8*, 475–489.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*, 1034–1037. https://doi.org/10.1126/science.1207998.

American Fedration of Teachers (2015). *Quality of worklife survey*. Retrieved from . August 10, 2017 ttps://www.aft.org/sites/default/files/worklifesurveyresults2015.pdf.

Arens, A. K., & Morin, A. J. S. (2016). Relations between teachers' emotional exhaustion and students' educational outcomes. *Journal of Educational Psychology, 108*(6), 800–813.

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in the classroom. *Journal of Applied Behavior Analysis, 2*, 119–124.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009a). Altering school climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10*, 100–115. https://doi.org/10.1007/s11121-008-0114-9.

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 Education & Psychology, 107*, 546–557.

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

Brown, C. H. (1993). Analyzing preventive trials with generalized additive models. *American Journal of Community Psychology, 21*, 635–664.

Browers, A., & Tomic, W. (2000). A longitudinal study of teacher burnout and perceived self-efficacy in classroom management. *Teaching and Teacher Education, 16*, 239–253.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673–682.

Debnam, K. J., Pas, E. T., Bottiani, J., Cash, A. H., & Bradshaw, C. P. (2015). An examination of the association between observed and self-reported culturally proficient teaching practices. *Psychology in the Schools, 52*, 533–548. https://doi.org/10.1002/pits.21845.

Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., ... Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology, 14*, 317–345.

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

Domitrovich, C. E., Bradshaw, C. P., Greenberg, M. T., Embry, D., Poduska, J. M., & Ialongo, N. S. (2009). Integrated preventive interventions: The theory and logic. *Psychology in the Schools, 47*, 71–88.

Downer, J. T., Pianta, R., Burchinal, M., Field, S., Hamre, B., LoCasale-Crouch, J., ... Scott-Little, C. (2018). *Coaching and coursework focused on teacher-child interactions during language/literacy instruction: Effects on teacher outcomes and children's classroom engagement.* (Unpublished manuscript available from the first author).

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer Quarterly (Wayne State University. Press), 55*, 224–249. https://doi.org/10.1353/mpq.0.0030.

Dumont, R., Willis, J. O., & Walrath, R. (2016). Clinical interpretation of the Woodcock-Johnson IV tests of cognitive abilities, academic achievement, and oral language. In R. Dumont, J. O. Willis, & R. Walrath (Eds.). *WJ IV clinical use and interpretation* (pp. 31–64). New York: Academic Press.

Ema Ushioda, E., & Dörnyei, Z. (2012). Motivation. In S. M. Gass, & A. Mackey (Eds.). *The Routledge handbook of second language acquisition* (pp. 396–409). New York: Routledge.

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

Evertson, C. M., & Weinstein, C. S. (Eds.). (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Mahwah, NJ: Lawrence Erlbaum Associates.

Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., & Vega, R. (2014). Effects of the good behavior game on challenging behaviors in school settings. *Review of Educational Research, 84*, 546–571. https://doi.org/10.3102/0034654314536781.

Greenberg, M., Weissberg, R., O Brien, M., Zins, J., Fredericks, L., Resnik, H., et al. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American Psychologist, 58*, 466–474.

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness, 10*, 40–67.

Greenberg, M. T., Brown, J. L., & Abenavoli, R. M. (2016). Teacher stress and health effects on teachers, students, and schools. [issue brief]. Retrieved from http://www.rwjf.org/content/dam/farm/reports/issuebriefs/2016/rwjf430428.

Hamre, B. K., Pianta, R. C., Burchinal, M., Field, S., Locasale-Crouch, J., Downer, J. T., ... Scott-Little, C. (2012a). Supporting effective teacher-child interactions through coursework: Effects on teacher beliefs, knowledge, and observed practice. *American Educational Research Journal, 49*, 88–123. https://doi.org/10.3102/0002831211434596.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2012b). Promoting young children's social competence through the preschool PATHS curriculum and MyTeachingPartner professional development resources. *Early Education and Development, 23*, 809–832.

Harmsen, R., Lorenz, M., Maulana, R., & van Veen, K. (2018). The relationship between beginning teachers' stress causes, stress responses, teaching behaviour and attrition. *Teachers and Teaching: Theory and Practice, 24*, 626–643. https://doi.org/10.1080/13540602.2018.1465404.

Hoglund, W. L., Klingle, K. E., & Hosan, N. E. (2015). Classroom risks and resources: Teacher burnout, classroom quality and children's adjustment in high needs elementary schools. *Journal of School Psychology, 53*(5), 337–357. https://doi.org/10.1016/j.jsp.2015.06.002.

Ialongo, N., Rogosch, F. A., Cicchetti, D., Toth, S. L., Buckley, J., Petras, H., & Neiderhiser, J. (2006). A developmental psychopathology approach to the prevention of mental health disorders. In D. Cicchetti, & D. J. Cohen (Eds.). *Developmental psychopathology* (pp. 968–1018). New York, NY: Wiley.

Ialongo, N., Werthamer, L., Brown, C. H., Kellam, S., & Wai, S. B. (1999). The proximal impact of two first grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology, 27*, 599–642.

Jennings, P., & Greenberg, M. (2009). The pro-social classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*, 491–525.

Kane, T., Rockoff, J., & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City. NBER working paper no. 12155*.

Kellam, S. G., Brown, C. H., Poduska, J., Ialongo, N., Wang, W., Toyinbo, P., … Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. Drug and Alcohol Dependence, 95(S1), S5–S28. doi:https://doi.org/10.1016/j.drugalcdep.2008.01.004.

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, 35*, 259–282.

Knapp, T. J. (1976). The Premack principle in human experimental and applied settings. *Behaviour Research and Therapy, 14*, 133–147. https://doi.org/10.1016/0005-7967(76)90067-X.

Luiselli, J. K., Putnam, R. F., Handler, M. W., & Feinberg, A. B. (2005). Whole-school positive behaviour support: Effects on student discipline problems and academic performance. *Educational Psychologist, 25*, 183–198.

Ma, X., & Willms, J. D. (2004). School disciplinary climate: Characteristics and effects on eight grade achievement. *Alberta Journal of Educational Research, 50*, 169–188.

Marzano, R. J., Gaddy, B. B., & Dean, C. (2000). *What works in classroom instruction.* Aurora, CO: Mid-continent Research for Education and Learning.

Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science, 14*, 179–196. https://doi.org/10.1080/10888691.2010.516187.

Maslach, C. M., Jackson, S. E., & Leiter, M. P. (1996). *Maslach burnout inventory: Manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.
McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*, 376–390.
Mikami, A. Y., Gregory, A., Allen, J. P., Pianta, R. C., & Lun, J. (2011). Effects of a teacher professional development intervention on peer relationships in secondary classrooms. *School Psychology Review, 40*, 352–366.
Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94*, 423–432.
Musci, R. J., Pas, E. T., Bettencourt, A. F., Masyn, K. E., Ialongo, N. S., & Bradshaw, C. P. (2019). How do student behavior and other classroom contextual factors relate to teachers' implementation of an evidence-based intervention? A multilevel SEM. *Development and Psychopathology, 31*, 1827–1835. https://doi.org/10.1017/S095457941900097X.
Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316.
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th edition). Los Angeles, CA: Muthén & Muthén.
National Institute for Occupational Safety and Health (1999). *Stress at work*. Washington, DC: U.S Printing Office.
Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*, 532–566.
Oliver, R., Wehby, J., & Daniel, J. (2011). Teacher classroom management practices: Effects on disruptive or aggressive student behavior. *Campbell Systematic Reviews, 2011*(4), https://doi.org/10.4073/csr.2011.4.
Oliver, R. M., & Reschly, D. J. (2007). *Effective classroom management: Teacher preparation and professional development. (TQ connection issue paper)*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved June 12, 2011, from http://www.tqsource.org/topics/effectiveClassroomManagement.pd.
Pas, E., Cash, A., O'Brennan, L., Debnam, K., & Bradshaw, C. P. (2015). Profiles of high school student behavior in the classroom: Associations with teacher behavior management strategies and classroom composition. *Journal of School Psychology, 53*, 137–148. https://doi.org/10.1016/j.jsp.2014.12.005.
Petras, H., Masyn, K., & Ialongo, N. (2011). The developmental impact of two first grade preventive interventions on aggressive/disruptive behavior in childhood and adolescence: An application of latent transition growth mixture modeling. *Prevention Science, 12*, 300–313. https://doi.org/10.1007/s11121-011-0216-7.
Pianta, R., La Paro, K., Payne, C., Cox, M., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102*, 225–238.
Pianta, R. C., Belsky, J., Houts, R., & Morrison, F. (2007). Opportunities to learn in America's elementary classrooms. *Science, 315*, 1795–1796. https://doi.org/10.1126/science.1139719.
Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher–child interactions in pre-kindergarten classrooms. *Early Child Research Quarterly, 23*, 431–451. https://doi.org/10.1016/j.ecresq.2008.02.001.
Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*, 437–448.
Rimm-Kaufman, S. E., & Hamre, B. (2010). The role of psychological and developmental science in efforts to improve teacher quality. *Teachers College Record, 112*, 2988–3023.
Rusby, J. C., Crowley, R., Sprague, J., & Biglan, A. (2011). Observations of the middle school environment: The context for student behavior beyond the classroom. *Psychology in the Schools, 48*, 400–415.
Sugai, G., Horner, R., & Gresham, F. (2001). Behaviorally effective school environments. In M. Shinn, G. Slover, & H. Walker (Eds.). *Interventions for academic and behavior problems*. Silver Spring, MD: National Association of School Psychologists.
Thomas, D. E., Bierman, K. L., & The Conduct Problems Prevention Research Group (2006). The impact of classroom aggression on the development of aggressive behavior problems in children. *Development and Psychopathology, 18*, 471–487. https://doi.org/10.1017/S0954579406060251.
Tomlinson, C. A., & Imbeau, M. (2010). *Leading and managing a differentiated classroom.* New York, NY: Houghton Mifflin.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Examiner's manual. Woodcock-Johnson III tests of achievement.* Itasca, IL: Riverside Publishing.