A State-wide Quasi-Experimental Effectiveness Study of the Scale-up of School-Wide Positive Behavioral Interventions and Supports

Elise T. Pas[1]

Ji Hoon Ryoo[2]

Rashelle Musci[1]

Catherine P. Bradshaw[3]

[1]Johns Hopkins University, Bloomberg School of Public Health
[2]University of Southern California, Keck School of Medicine
[3]University of Virginia, Curry School of Education

Corresponding Author: Elise T. Pas; Johns Hopkins University, Bloomberg School of Public Health; 415 N. Washington Street, Office 507; Baltimore, MD 21231, USA; epas1@jhu.edu

Author Contacts:
Ji Hoon Ryoo; University of Southern California, Keck School of Medicine; 4640 West Sunset Boulevard; Los Angeles, CA 90033, USA; jryoo@ usc.edu

Rashelle Musci; Johns Hopkins University, Bloomberg School of Public Health; 624 N. Baltimore Street, Office 803; Baltimore, MD 21205, USA; rmusci1@jhu.edu

Catherine P. Bradshaw; University of Virginia, Curry School of Education; 112D Bavaro Hall Charlottesville, VA 22904, USA; Catherine.Bradshaw@ virginia.edu

Declarations of interest: None

This is a post-peer-review, pre-copyedit version of an article published in Journal of School Psychology. The final authenticated version is available online at: https://doi.org/10.1016/j.jsp.2019.03.001

Pas, E.T., Ryoo, J.H., Musci, R.J., and Bradshaw, C. P. (2019) A state-wide quasi-experimental effectiveness study of the scale-up of school-wide Positive Behavioral Interventions and Supports. *Journal of School Psychology*, ISSN: 0022-4405, Vol: 73, Page: 41-55 April 2019.
.
Agency: Institute of Education Sciences Grant Number: R305H150027

Abstract

The three-tiered Positive Behavioral Interventions and Supports (PBIS) framework promotes the development of systems and data analysis to guide the selection and implementation of evidence-based practices across multiple tiers. The current study examined the effects of universal (tier 1) or school-wide PBIS (SW-PBIS) in one state's scale-up of this tier of the framework. Annual propensity score weights were generated to examine the longitudinal effects of SW-PBIS from 2006-07 through 2011-12. School-level archival and administrative data outcomes were examined using panel models with an autoregressive structure. The sample included 1,316 elementary, middle, and high schools. Elementary schools trained in SW-PBIS demonstrated statistically significantly lower suspensions during the fourth and fifth study years (i.e., small effect size) and higher reading and math proficiency rates during the first two study years as well as in one and two later years (i.e., small to large effect sizes), respectively. Secondary schools implementing SW-PBIS had statistically significantly lower suspensions and truancy rates during the second study year and higher reading and math proficiency rates during the second and third study years. These findings demonstrate medium effect sizes for all outcomes except suspensions. Given the widespread use of SW-PBIS across nearly 26,000 schools in the U.S., this study has important implications for educational practices and policies.

Keywords: Positive Behavioral Interventions and Supports (PBIS); dissemination; state-wide effects; propensity score weighting

A State-wide Quasi-Experimental Effectiveness Study of the Scale-up of School-Wide Positive Behavioral Interventions and Supports

Positive Behavioral Interventions and Supports (PBIS; Sugai & Horner, 2002, 2006; Sugai, Horner, & Gresham, 2002) is a school-based, multi-tiered prevention framework that integrates data to inform decisions about practices and systems needed in the school to promote positive student behavior. At the universal (tier 1) level, referred to as school-wide PBIS (or SW-PBIS), there is a focus on shifting all staff toward a proactive and positive approach to behavior management and ensuring consistent implementation across all school settings (i.e., classroom and non-classroom). As described in greater detail below, prior PBIS efficacy research has largely focused on SW-PBIS outcomes in elementary schools and has demonstrated significant effects (a) across a range of student behavioral, social emotional, and academic outcomes; (b) student need for additional supports; and (c) school climate (e.g., Bradshaw, Koth, Bevans, Ialongo, & Leaf, 2008; Bradshaw, Koth, Thornton, & Leaf, 2009; Bradshaw, Waasdorp, & Leaf, 2012; Horner et al., 2009). Although the evidence base for PBIS continues to grow, less is known about the effects of SW-PBIS in regular practice when scaled up by a state. Moreover, much of the prior SW-PBIS scale-up research has used correlational designs and lacked a comparison group. Yet, the issue of effectiveness is particularly salient, given that SW-PBIS has been widely disseminated to nearly 26,000 schools across the United States (Horner et al., 2014; Sugai, Horner, & McIntosh, 2016) and internationally. The current study aimed to fill this gap by examining the real-world outcomes of SW-PBIS when scaled-up across a state, using a quasi-experimental non-equivalent control group design (Shadish, Cook, & Campbell, 2001).

Theoretical Background for PBIS

PBIS is based on behavioral, social learning, and organizational behavioral principles which, taken together, suggest that shifting the school environment can shape student behavior in a positive way. As adults model positive behaviors, more students will engage in such positive behaviors. As mentioned above, PBIS is a three-tiered prevention framework, where a universal system of supports is integrated with targeted (tier 2) and intensive (tier 3) preventive interventions for students displaying a higher level of need (O'Connell, Boat, & Warner, 2009). Consistent with a public health approach, it is expected that 80% of students within the building will respond to this universal system of behavioral supports, and the data and systems will be used to identify the roughly 15% of students with a need for more targeted or group intervention and the 1-5% of students in need of individualized and intensive supports (O'Connell et al., 2009). This same tiered framework is commonly used to promote academic learning, whereby the universal curriculum and supports are provided to meet the needs of the majority of the students, and more intensive academic supports are provided at tiers 2 and 3 for students needing greater assistance to develop their skills (Arden, Gruner Gandhi, Zumeta Edmonds, & Danielson, 2017).

Core Components of SW-PBIS

Training in multi-tiered PBIS has a strong emphasis on data, systems, and practices across the intervention continuum. SW-PBIS training specifically focuses on data collection regarding implementation of core features of the model, data on behavioral infractions, as well as other data points that can be used as a means for assessing when students respond positively to the universal behavioral supports or may need additional targeted or intensive supports (Horner, Sugai, Todd, & Lewis-Palmer, 2005; Horner, Sugai, & Anderson, 2010). Systems that allow for consistent implementation and collection and analysis of data are needed. This, in turn, informs

data-based decision making, the selection and evaluation of discrete teacher practices, and the provision of on-going professional development that the SW-PBIS team provides to all school personnel. The intersection of data, systems, and practices would be expected to be mirrored in any advanced tier implementation efforts as well.

The current study focused on the scaled-up implementation of the universal, school-wide component of PBIS, which explicitly targets the school's systems and procedures to prevent and respond to disruptive behavior, with an emphasis on clarity and consistency. Through training in SW-PBIS, school staff learn to set and teach clear behavioral expectations, implement a system to respond to the meeting of behavioral expectations (i.e., as a means for proactively encouraging desired and preventing undesired behaviors), and create and implement a consistent response system to behavioral infractions for all students across all school settings (Sugai & Horner, 2002, 2006). In doing so, the expectation is that students will engage in fewer disruptions and receive fewer classroom removals, and thus will experience increased time for instruction and learning, which will translate into improved academic performance (Sugai et al., 2002). Thus, improvements in behavior are expected to be proximal outcomes and academic outcomes are expected to be more distal. At the time of data collection for the current study, training and support for the advanced tiers (i.e., 2 or 3) was not systematically or widely available within the state.

Prior Efficacy Research on SW-PBIS and Multi-tiered PBIS

A series of randomized controlled trials (RCTs) testing PBIS in elementary schools has provided an evidence base for its efficacy (also see Horner et al., 2010). Specifically, two RCTs conducted in elementary schools provide evidence that tier 1 SW-PBIS was associated with reduced student office discipline referrals and suspensions, improved school climate (Bradshaw,

Koth, et al., 2008; Bradshaw, Koth, et al., 2009; Bradshaw, Mitchell, & Leaf, 2010) and improved student academic achievement (Horner et al., 2009). More specifically, the RCT with papers authored by Bradshaw and colleagues demonstrated that the overall referral rate was reduced by approximately 18% in SW-PBIS schools and students in SW-PBIS schools were 33% less likely to receive a referral than students in comparison schools. Further, small- to medium-sized effects were evinced (i.e., *d*s of .10 to .30) on measures of climate. Schools implementing SW-PBIS also rated their students as needing fewer specialized support services (Bradshaw, Waasdorp, & Leaf, 2012) and as having fewer behavioral problems (e.g., aggressive behavior, concentration problems, bullying, rejection; Bradshaw, Waasdorp, et al., 2012; Waasdorp, Bradshaw, & Leaf, 2012). The effect sizes for these outcomes also were in the small range. A generalizability study (Stuart, Bradshaw, & Leaf, 2015) leveraged data from this Maryland-based RCT and demonstrated that the positive effects generalized when schools in the trial were weighted to match the characteristics of schools within the state.

A third elementary school RCT involved schools all trained in SW-PBIS and the intervention schools further incorporated an external coach to provide tailored training and implementation support for the student support teaming process and for the implementation of targeted behavioral and engagement interventions (i.e., Tier 2). The aim of this RCT was to examine the effects of multi-tiered PBIS. In this 42-school RCT, teachers in the intervention schools reported small improvements in student need for special education, student academic performance, and their own self-efficacy (see Bradshaw, Pas, Goldweber, Rosenberg, & Leaf, 2012).

In a fourth RCT, in a high school, an external coach assisted in the integration of school climate survey data into the data-based decision-making of PBIS. The coach also offered training

and on-going supports in evidence-based programs targeting the universal prevention of bullying or substance use and targeted interventions to improve student engagement or experiences of trauma. Student surveys regarding safety (i.e., weapon carrying, being threatened to be injured with a weapon, skipping school because of a fear of safety) and overall engagement across multiple domains improved by the end of the first year of implementation (see Bradshaw, Debnam, et al., 2014).

Dissemination and Implementation Research on SW-PBIS

State-wide program evaluations of SW-PBIS effectiveness have generally shown promising findings, indicating trends of lower office discipline referrals and suspensions in implementing schools (i.e., no comparison group; Barrett, Bradshaw, & Lewis-Palmer, 2008; Childs, Kincaid, George, & Gage, 2016; Freeman et al., 2016; Muscott, Mann, & LeBrun, 2008). In Maryland, studies regarding the scale-up have most recently focused on the dissemination process and implementation fidelity rather than effectiveness. Specifically, in examining the characteristics of schools and PBIS training and adoption, findings indicated that schools with more suspensions were more likely to be trained in PBIS and schools with greater student mobility and poorer student academic proficiency were more likely to be trained in and to adopt PBIS (i.e., implement and submit implementation data to the state consortium; Bradshaw & Pas, 2011). PBIS implementation fidelity scores were highest in schools that had (a) implemented for a greater number of years and (b) had more certified teachers working in the building, as measured by the Implementation Phases Inventory (i.e., IPI; Bradshaw, Debnam, Koth, & Leaf, 2009; Bradshaw & Pas, 2011). Scores on the IPI were also associated with school-level student outcomes in elementary and middle schools, such that higher IPI scores were associated with

higher academic proficiency rates on state standardized math and reading assessments as well as lower rates of truancy (Pas & Bradshaw, 2012).

To our knowledge, there have been two dissemination studies using methodological approaches that have taken steps toward drawing causal inferences (e.g., by minimizing threats to validity such as selection bias) about the effectiveness of SW-PBIS when disseminated within a state in conjunction with district partners. The first was a study conducted in Minnesota among a relatively small sample of trained schools (i.e., 32 elementary and 34 middle schools; Ryoo, Hong, Bart, Shin, & Bradshaw, 2018). A second recent study was conducted across the state of Florida, matching schools implementing SW-PBIS with fidelity with those never trained in SW-PBIS (Gage, Grasley-Boy, George, Childs, & Kincaid, 2019). The Florida study demonstrated that schools implementing SW-PBIS with fidelity had lower suspension rates than non-PBIS schools. However, the Florida study focused solely on one year's data and only examined school discipline outcomes. Thus, the current study fills important gaps in extant literature regarding the effects of SW-PBIS when scaled-up throughout a state and across a wide range of high stakes student outcomes.

Training in and Scaling of SW-PBIS

Training for PBIS implementation in the United States is provided by the federal Office for Special Education Programs, and the costs of implementing PBIS are relatively low (Horner et al., 2012), which may explain its expansive scaling. Nearly all states in the United States have developed a state- or district-level infrastructure to support its implementation; several other countries are also scaling up PBIS (e.g., Canada, Australia). In Maryland, where the current study was conducted, a coordinated system for implementation of SW-PBIS has been developed over nearly two decades, through collaboration between the Maryland State Department of

Education, Sheppard Pratt Health System, and Johns Hopkins University (Barrett et al., 2008; Bradshaw, Debnam, et al., 2014; Bradshaw & Pas, 2011), or the state management team. This collaborative, called the PBIS Maryland Consortium, also has a state leadership team with a representative from each of these agencies as well as from the 24 local education agencies (i.e., school districts) in the state. There is ongoing data collection and evaluation of implementation and outcomes by the state management team (for details, see Barrett et al., 2008; Bradshaw, Debnam, et al., 2014). During the time frame of this study, there were annual, two-day state-wide offerings of initial SW-PBIS trainings for new teams and booster trainings for returning teams, quarterly full-day state leadership meetings to train district contacts and ensure that state-wide trainings were aligned to their needs, and quarterly full-day SW-PBIS coaches trainings provided to school-based PBIS coaches throughout each school year; all training efforts were led by the PBIS state-level management team (see Barrett et al., 2008). School-based coaches and district leaders (Rogers, 2002; Schoenwald & Hoagwood, 2001) help to promote fidelity and on-going implementation of SW-PBIS. For example, districts offer their own monthly or quarterly coaches' meetings for additional professional development support.

In total, there are currently about 1,100 Maryland schools (i.e., pre-k, elementary, middle, high, alternative, special education) trained in and 855 schools actively implementing SW-PBIS and providing data to the statewide collaborative. The state is now beginning to disseminate training and webinars about implementation of PBIS at the more advanced (i.e., targeted and intensive) tiers for students not responding to SW-PBIS. The data regarding the training status and implementation levels for the current study come from the state's evaluation efforts.

Overview of the Current Study

Taken together, extant efficacy research has suggested significant positive effects of PBIS on a range of behavioral and academic outcomes. There has been less consideration of the effectiveness of PBIS within the context of state-wide scaling; however, a recent state-wide study in Florida examined discipline outcomes and reported effects on suspensions (Gage et al., 2019). Most of the available scale-up studies have lacked comparison groups and suffer from threats to validity, including selection bias (Barrett et al., 2008; Bradshaw & Pas, 2011; Bradshaw, Pas, Barrett, et al., 2012; Childs et al., 2016; Freeman et al., 2016). Additional rigorous research that takes steps toward eliminating such threats to validity, and gets closer to drawing causal inferences about the impacts of PBIS when widely disseminated is needed. The current study was designed to fill this important gap in the effectiveness research on PBIS by examining the effectiveness of PBIS on a range of student outcomes when scaled-up within the state of Maryland. Our first aim was to examine the levels of implementation achieved among SW-PBIS schools as a means for confirming that training status in SW-PBIS did in fact lead to school-based implementation and for contextualizing what "regular practice" in Maryland is (i.e., whether adequate fidelity was the norm within the state). Our second aim was to determine whether training in SW-PBIS was associated with improved student outcomes.

A quasi-experimental non-equivalent control group design (Shadish et al., 2001) was selected, in which we leveraged existing archival data and used propensity score weights to approximate a control condition comprised of non-trained schools (Rosenbaum & Rubin, 1983). In other words, although there are differences between schools that have selected to be trained and not selected to be trained in SW-PBIS, these differences can be measured and observed. By accounting for the differences in these observed variables and the likelihood that any school would be in the intervention group, we can then weight the data from each school to either

contribute more or less information in the outcome analysis. Using propensity score weights also allows for all schools in the state to remain in the outcome analysis; other approaches would result in the dropping of schools that were too dissimilar from other schools, therefore biasing the sample. The data for this study came from the state-wide scale-up and evaluation of SW-PBIS in Maryland public schools, as implemented by existing school personnel. We focused on PBIS training and implementation which occurred in 2006-07 through 2011-12 among public elementary and secondary schools. We hypothesized that schools trained in SW-PBIS would demonstrate lower rates of suspensions and truancy and higher levels of academic proficiency, based both on the findings of prior RCTs (Bradshaw et al., 2010; Bradshaw, Waasdorp, et al., 2012; Horner et al., 2009) and non-experimental dissemination studies (Bradshaw & Pas, 2011; Bradshaw, Pas, Barrett, et al., 2012; Childs et al., 2016; Freeman et al., 2016). Based on the conceptual model for change, we also hypothesized that improvements in suspensions may emerge in the earlier years, as this is the most proximal outcome for PBIS, whereas the truancy and academic effects would emerge later.

Method

Participants

Eligibility. Within the state of Maryland, there are 24 districts or local education agencies (i.e., 23 counties and one city), all of which have some schools that participated in the Maryland SW-PBIS Initiative. The focus of this study was on traditional elementary, middle, and high schools (i.e., settings only for students receiving special education and alternative schools were excluded). Elementary schools included K-5 or K-6 as well as K-8 schools (referred to from here on as elementary schools); secondary schools included traditional middle schools (grades 6-8), traditional high schools (grades 9-12), and combined middle/high schools (i.e.,

grades 6-12). There were 1,316 schools across the 24 districts that qualified as defined above and were open during the study time frame, of which 859 were trained in SW-PBIS and 457 were never trained. Elementary schools comprised 67% of the sample (i.e., n = 879) and secondary (i.e., middle and high) schools comprised 33% of the sample (i.e., n = 437).

The schools in Maryland are, on average, large schools (i.e., with an average enrollment of over 600 students, ranging from 635.46 to 651.89 across study years) serving a diverse student body. Specifically, White students comprised the majority of the sample (41.75% to 47.51%), followed by African American students (36.29% to 38.36%), and Hispanic students (8.64% to 11.61% in study years). Over the course of the study, there was a decrease in the proportion of White students and increase in Hispanic students. Asian students pretty consistently comprised about 5% of the sample and less than 1% of students were American Indian/Native American. With regard to the outcomes across the entire sample, the suspensions rate declined steadily from 11.05% in 2006-07 to 8.22% in 2011-12. Truancy rates ranged from 8.47 to 9.72% and academic proficiency rates were below 80% in 2006-07 and above 80% in all subsequent years. See Table 1 for all unweighted demographic averages broken out by school level and PBIS status in 2006-07.

Training Procedures

As noted earlier, all training was offered through state-wide SW-PBIS training opportunities. These initial and booster trainings were two-day trainings that required schools to gain 80% buy-in from staff members to implement SW-PBIS, make a three-year commitment to implementing, and identify a 4-6 person team including an administrator and coach who would attend training (see Bradshaw & Pas, 2011). The on-going support to schools from the state was directed at the team (annually) and the coach (quarterly); the team was expected to provide the

schools with on-going supports (e.g., creating the vision and materials, providing training). The school-level implementation was tracked through state-wide data collection (see below).

Measures

Training Status and Implementation. Personnel from the Sheppard Pratt Health System (SPHS) and the Maryland State Department of Education served as implementation partners and provided trainings throughout the state; the partnership also collected data regarding the year in which schools were trained and implementation status and fidelity over time. These data were shared with the university-based research partners for the current analysis and were approved by the relevant Institutional Review Boards. The year in which a school was trained was provided and recoded to training status (0 = not trained, 1 = trained). For the current analysis, once a school was considered trained, it could not be returned to an untrained status. See Table 2 for annual training data.

Implementation data were collected during the fall and spring of each school year and served as an indicator of implementation fidelity for the schools across the state. Specifically, each spring, schools submitted the School-Wide Evaluation Tool (SET) measure (Sugai, Lewis-Palmer, Todd, & Horner, 2001) as part of the process to receive recognition for PBIS implementation. The SET was the most widely-used measure of the core features of the universal SW-PBIS model during this time and has been included in extant efficacy research. Previous studies have documented that it is reliable and valid (Bradshaw, Reinke, Brown, Bevans, & Leaf, 2008; Horner et al., 2004). Internal consistency has been demonstrated across a series of RCTs and other studies (e.g., Pas, Johnson, Debnam, Hulleman, & Bradshaw, 2019); the reported alphas below were from research conducted in 198 elementary, middle, and high schools. The

SET consists of seven subscales that assess the degree to which schools implement the key features of SW-PBIS (Horner et al., 2004). The seven included scales are: (a) *Expectations Defined* (2 items; Cronbach's alpha [α] = .78)*;* (b) *Behavioral Expectations Taught* (5 items; α = .90); (c) *System for Rewarding Behavioral Expectations* (3 items; α = .86); (d) *System for Responding to Behavioral Violations* (5 items; α = .51); (e) *Monitoring and Evaluation* (4 items; α = .79); (f) *Management* (8 items; α = .92); and (g) *District-Level Support* (2 items; α = .55). Each item of the SET is scored on a 3-point scale from 0 (*not implemented*) to 2 (*fully implemented*), and a scale score reflecting the percentage of earned points is calculated. Higher scores reflect greater implementation fidelity. The scores on all scales were averaged to calculate one total score. Within Maryland, a district representative, state personnel, or university contractor administered the SET through a half-day site visit, during which brief interviews were performed with school leadership, staff, and students; documents were also reviewed and observations conducted as further evidence of implementation of SW-PBIS.

Bi-annually (i.e., in the fall and spring of each year), schools also submitted data on the Implementation Phases Inventory (IPI; Bradshaw, Debnam, et al., 2009). This measure was used by the state as an indicator of active implementation status and is unique in that it follows a "stages of change" theoretical model, thereby capturing which of a series of four stages that the school has reached: *preparation* (α =.65, e.g., "PBIS team has been established", "School has a coach"), *initiation* (α = .80, e.g., "A strategy for collecting discipline data has been developed", "New personnel have been oriented to PBIS"), *implementation* (α = .90, e.g., "Discipline data are summarized and reported to staff", "PBIS team uses data to make suggestions regarding PBIS implementation"), and *maintenance* (α = .91, e.g., "A set of materials has been developed to sustain PBIS", "Parents are involved in PBIS related activities"). In total, there are 44 items

assessing these key elements of SW-PBIS. This measure was completed by the PBIS coach, who indicated the level of implementation for each element on a 3-point scale from 0 (*not in place*) to 2 (*fully in place*). The percentage of implemented elements was calculated for each stage, such that a higher score indicates greater implementation. The scores on these four stages were averaged to calculate one total score for this study. The IPI incorporates a broader set of implementation components, provides a different (i.e., school personnel) lens on implementation, and has demonstrated fewer ceiling effects. The IPI was included in this study as a second indicator of implementation fidelity given its prior demonstrated association with student outcomes (Pas & Bradshaw, 2012) and because it allowed for a more inclusive and broader set of fidelity indicators and larger sample of schools with implementation data within the state. A previous study of the psychometric properties of the IPI found it to have adequate internal consistency ($\alpha = .94$) and a test-retest correlation of .80 (Bradshaw, Debnam, et al., 2009). See Table 2 for annual spring SET and IPI scores throughout the time frame of this study.

School-level outcomes. The school outcome data were provided by the Maryland State Department of Education (MSDE). These include the (a) school-level suspension rates (i.e., total suspension events divided by total school enrollment times 100; i.e., not the percent of students suspended.); (b) truancy rates (i.e., percent of students missing 20 or more days of school across a given school year); and (c) percent of students within each school who were proficient on tests of academic (i.e., reading and math) proficiency. The standardized achievement assessments varied based on school level, but were consistent across time in structure and administration. Specifically, the Maryland School Assessments for reading and mathematics were completed by grades 3-5 in elementary schools and all grades (i.e., 6-8) in middle schools. The percent of students in each given year who attained a proficient or advanced score on the English 2 and

Algebra High School Assessment (i.e., HSA) were utilized for high schools. The grade levels in which these tests are taken vary, based on when a student completes the course, but the on-time course completion is 10th grade for English 2 and 9th for Algebra. School rates were calculated by averaging the percent of students who were proficient and advanced in each assessed grade or subject area. The outcome data included data from 2006-07 through 2011-12; data on these indicators from 2004-05 were also included for the propensity score weights. Average baseline rates and difference scores throughout the time frame of this study are depicted in Table 3.

School-level demographic characteristics. The demographic information regarding the schools throughout the state was also provided by MSDE. Demographics from 2006-07 (i.e., the first year of the study) were included in the propensity score weighting and outcome analyses. Data regarding (a) student enrollment (i.e., the number of students in the school), (b) student mobility (i.e., percent of students who entered the school, plus the percentage who withdrew from the school, divided by total student enrollment), and the (c) percent of students receiving free and reduced-priced meals were utilized for the propensity score weighting and were also controlled for in the outcome analyses. The percent of students in each racial/ethnic group was also considered and included these five groups: White, American Indian/Native America, Asian, Hispanic, and African American. Each group was dummy coded and included in the propensity score weighting; only White (i.e., versus all others) was used in the outcome analyses. Additional data were only considered for the propensity score weighting, including the percent of students receiving special education and English language services and the student-teacher ratio. These variables are included because: (a) prior research demonstrated that such demographic data were associated with being trained in SW-PBIS and subsequently submitting data (Bradshaw & Pas, 2011; Pas & Bradshaw, 2012; Ryoo et al., 2018) and thus are considered confounders,

representing selection bias, for treatement status; (b) correlational analyses demonstrated significant associations between the demographic variables, years since training, and implementation levels in this study; and (c) similar variables were used in a prior generalizability (Stuart et al., 2015) and propensity score study (Ryoo et al., 2018). Table 1 displays the unweighted means for each of these demographic variables for schools trained and not trained in SW-PBIS. Because all outcome data were state collected, data were consistently present within a given year and no more 7% of data were missing in a given year; missing data was due to schools not operating in a given year.

Analyses

Descriptive analyses (i.e., means, standard deviations) were used to summarize the annual implementation fidelity on the SET and IPI measures across the time frame of this study. For the outcome analyses, the goal was to include all elementary and secondary (i.e., middle and high) schools across the state; however, training in SW-PBIS was a choice made by schools and thus was not a controlled variable. Therefore, we minimized the effect of the possible selection biases by applying propensity score methods (i.e., PSMs; Rosenbaum & Rubin, 1983) which allows for all schools to remain in the sample, but balances the baseline differences by allowing for some schools to provide more information to the analysis than others.

Among the choices for PSMs are matching, subclassification, and weighting. We conducted propensity score weighting (PSW; Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; Rosenbaum, 1987) in the *R software* using the Twang package (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2016) to reduce the selection biases. The weights were calculated using the average treatment effect for the treated (ATT; McCaffrey et al., 2013) because our interest was whether SW-PBIS was beneficial for SW-PBIS schools, assuming that

the comparison schools were also providing programming related to student behavior (Winship & Morgan, 1999). The weights using the ATT also addressed the time-varying nature of PBIS status, whereby schools may have changed from untrained to trained in PBIS during the study time frame.

The core set of variables included in the PSW modeling each year included outcomes of interest in 2004-05 (i.e., suspensions, truancy, and reading and math proficiency) as well as 2006-07 data for enrollment and the percent of students who were in each of the racial/ethnic groups (i.e., American Indian/Native, Asian, Hispanic, African American, and White) and received free and reduced-priced meals. Additional variables were added to the PSWs, incrementally, to ensure improvements in balance and not adding redundancy. The additional variables included the (a) percent of students receiving special education, (b) percent of students receiving English language services, (c) mobility rate, and (d) student-teacher ratio. In the final PSW models, all of these listed variables, except for percent of students receiving English language services, were included, as this model demonstrated the most consistent and best balance across subsamples.

To examine whether schools trained in SW-PBIS had better suspension, truancy, and reading and math proficiency rates across six years than non-trained schools, a series of panel models with an autoregressive structure (Kline, 2016) was conducted in the M*plus* software (Muthén & Muthén, 2002-2018). Annual difference scores served as the outcomes. For example, the 2007-08 suspension outcome reflected the difference between suspensions in 2007-08, as compared to 2006-07. Such difference scores allowed us to discretely examine changes to outcomes in specific years, as opposed to modeling one slope estimate for the entire time frame, as would be employed in a growth mixture modeling approach (Little, 2013). This analytic

approach is important, as effects may be time sensitive (e.g., emerging early or emerging later). The annual changes in each of the four outcome variables were modeled utilizing the year-specific propensity score weights. Therefore, all schools were included in the models, just with varying weight during each year and the multiple (i.e., six) years of data were accounted for.

Annual PBIS status variables (for which a school's status could change from comparison (0) to intervention (1) over time) were the independent variables of interest. The models also controlled for all other 2004-05 outcome values (except math and reading were not included in the same model because of collinearity), as well as additional covariates collected in 2006-07 (enrollment, free and reduced-priced meals, mobility, and percent of students in the building who were White; see Figure 1).

The model equation for the set of repeated measures on outcome y is

$$y_{it} = \alpha_i + \sum_{j=1}^{5} \lambda_{tj}\beta_{ij} + \sum_{k=1}^{t-1} \rho_{t,t-k} y_{i,t-k} + \varepsilon_{it},$$

where $t = 2,\ldots,6$, $E(\varepsilon_{it}) = 0$, $Cov(\varepsilon_{it}, y_{i,t-k}) = 0$, $Cov(\varepsilon_{it}, \beta_{ij}) = 0$, and $Cov(\varepsilon_{it}, \alpha_i) = 0$; α_i = the intercept for the outcome y_{it}; λ_{tj} = the regression coefficient of a covariate β_{ij}; and $\rho_{t,t-k}$ = the regression coefficient of a prior outcome, $y_{i,t-k}$. We further assume that $Cov(\varepsilon_{it}, \varepsilon_{lt}) = 0$ for all t and $i \neq l$, $Var(\varepsilon_{it}) = \sigma^2_{\varepsilon_i}$ for each t, and $Cov(\varepsilon_{it}, \varepsilon_{i,t+m}) = 0$ for $m \neq 0$.

All of the models were evaluated with the root mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean square residual (SRMR). The criteria for acceptable model fit are less than 0.08 for both RMSEA and SRMR, and greater than 0.90 for CFI (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999). We also calculated Cohen's *d* effect sizes (Cohen, 1988) by subtracting the weighted mean differences of

each outcome for the trained and untrained schools and dividing by the weighted pooled standard deviation. Effect sizes for statistically significant SW-PBIS effects are reported in text.

Results

Implementation Levels

In elementary schools, the average scores on the SET and IPI exceeded the 80% benchmark in all years. In fact, SET scores were on average over 90% in all but the first year, and the IPI averages ranged from 83.5% to 90.2%. SET scores had low standard deviations in all but the first year and the majority of schools achieved high fidelity on the SET measure. In secondary schools, the average scores on the SET exceeded the 80% benchmark in all years, whereas the IPI scores exceeded 80%, on average, in 2008-09 and beyond. As was observed in the elementary schools, SET scores in secondary schools were on average over 90% in all but the first year, and generally had low standard deviations. In other words, trained schools in this study, on average, demonstrated adequate to high fidelity.

Balancing Data Using Propensity Score Weighting Method

Applying generalized boosted modeling (GBM; McCaffrey, Ridgeway, & Morral, 2004), we estimated propensity score weights for yearly datasets from the 2006-07 school year through the 2011-12 school year, using school-level variables. Using these weights, we conducted balance checks before running outcome analyses, which not only indicated how well the propensity score weighting method reduced selection biases, but also are useful for describing the results of the causal analyses. See Table 1 for a listing of weighted and unweighted means and the effect sizes, as demonstrated by the standardized mean differences between groups.

Figure 2 contains plots for assessing the balance between groups on the trained in SW-PBIS variable before and after weighting for elementary schools in 2006-07 (i.e., first year of the

outcome analyses). Included in these plots are the standardized effect size, which are defined as the "treatment group mean minus the control group mean divided by the treatment group mean" (Ridgeway et al., 2016, p. 8). Applying the criteria that standardized mean differences of less than 0.20 are considered 'small', 0.40 are considered 'moderate', and 0.60 are considered 'large' (Cohen, 1988; Ridgeway et al., 2016), we confirmed that the 'moderate' or 'large' differences before propensity score weighting were reduced to 'small' for all variables; the minor exception was the baseline suspension rate at 2004-05 within secondary schools was reduced to 0.21 (i.e., 0.01 above small; See Table 1). The balance tables along with figures for other years also indicated that propensity score weighting balanced the data between SW-PBIS and non-SW-PBIS schools over the study years from 2006-07 to 2011-12. For each outcome year, prior year's data was controlled for to ensure that any remaining differences were fully accounted for.

Findings for Elementary Schools

Model fit. In all four of the elementary models, RMSEA (0.000 for suspension with 90% CI [0.000, 0.024], 0.033 for truancy with 90% CI [0.019, 0.046], 0.038 for reading with 90% CI [0.026, 0.051], and 0.037 for math with 90% CI [0.024, 0.050]), CFI (1.000 for suspension, 0.926 for truancy, 0.927 for reading, and 0.912 for math) and SRMR (0.009 for suspension, 0.011 for truancy, 0.013 for reading, and 0.012 for math) were within the acceptable ranges.

SW-PBIS effects. In elementary schools, there was a significant effect of SW-PBIS on the suspension difference scores in 2009-10 and 2010-11 (d = 0.17 and 0.18, respectively). These effect sizes correspond to about a 1% suspension rate improvement for SW-PBIS elementary schools in each of these years. Although suspension rates generally improved for all schools across this time period, the reduction in rates for suspensions in SW-PBIS elementary schools were statistically greater than those in non-trained schools during 2009-10 and 2010-11. With

regard to truancy, there were no statistically significant changes for elementary schools trained in SW-PBIS (see Table 4 for full listing of results).

Scores on the elementary reading and mathematics assessments (i.e., MSA) generally improved for all schools during this time frame (see Table 3). The elementary schools trained in SW-PBIS demonstrated statistically significant increases in reading proficiency in 2006-07 (d = 0.32), 2007-08 (d = 1.00), and 2010-11 (d = 0.30), as compared to non-trained elementary schools. These reflect improvements of 1.4% to 5% more students proficient in reading in SW-PBIS elementary than non-PBIS elementary schools. These same statistically significant findings emerged for mathematics proficiency in 2006-07 (d = 0.63), 2007-08 (d = 0.34), 2009-10 (d = 0.31), and in 2011-12 (d = 0.23). These reflect improvements of 1 to 4% more students proficient in math. All reported results are for models including the four school-level covariates from the 2006-07 school year (i.e., enrollment, free and reduced-priced meals, mobility, and percent of students in the building who were White), as well as the prior years' outcomes.

Findings for Secondary Schools

Model fit. In all four of the secondary models, the RMSEA (0.000 for suspension with 90% CI [0.000, 0.031], 0.031 for truancy with 90% CI [0.000, 0.052], 0.018 for reading with 90% CI [0.000, 0.044], and 0.033 for math with 90% CI [0.000, 0.055]), SRMR (0.010 for suspension, 0.016 for truancy, 0.015 for reading, and 0.015 for math), and CFI (1.000 for suspension, 0.950 for truancy, 0.989 for reading, and 0.942 for math) were within the acceptable ranges.

SW-PBIS effects. In secondary schools, we found positive and statistically significant effects of SW-PBIS on all four outcomes in 2007-08, as well as for reading and math proficiency in 2008-09. Specifically, SW-PBIS schools showed greater declines in suspensions and the

truancy rate and greater improvements in math and reading proficiency (i.e., on the MSA and HSA) during these two years. The effect sizes for these findings ranged from small to medium. In 2007-08, the effect sizes were 0.03 for suspensions (i.e., reflecting a less than half-percent improvement in the suspension rate), 0.43 for truancy (i.e., reflecting an improvement of 1.7% in truancy), 0.58 for reading (i.e., reflecting an improvement of 9% students proficient in reading), and 0.46 for math (i.e., reflecting an improvement of 8% students proficient in math). In 2008-09, the effect sizes were 0.53 for reading (i.e., reflecting an improvement of 1.9% students proficient in reading) and 0.30 for math (i.e., reflecting an improvement of 1.2% students proficient in math). There were no statistically significant differences in the changes in these outcomes between the trained and non-trained secondary schools in the other four study years (see Table 4).

Discussion

The purpose of this study was to examine the effectiveness of the state-wide scale-up of SW-PBIS at improving school-level behavioral outcomes and academic proficiency. This quasi-experimental non-equivalent control group design allowed us to remove selection biases that most extant literature has not addressed, and examined the effects of SW-PBIS across an entire state when translated into broad-scale practice through state infrastructure. This study fills an important gap in the extant literature, which had previously documented positive effects of SW-PBIS across myriad student behavioral and academic outcomes when studied in randomized controlled trials, mostly focused on elementary schools (e.g., Bradshaw, Koth, et al., 2008; Bradshaw, Koth, et al., 2009; Bradshaw, Mitchell, & Leaf, 2010; Bradshaw, Waasdorp, et al., 2012; Horner et al., 2009; Waasdorp, Bradshaw, & Leaf, 2012; also see Horner et al., 2010). Similarly, prior scale-up and dissemination studies of SW-PBIS have suggested that SW-PBIS is

positively associated with improved behavior outcomes; however, these studies with the exception of two recent studies (i.e., Gage et al., 2019; Ryoo et al., 2018) have been conducted without a comparison group (e.g., using pre- and post-test designs; see Barrett et al., 2008; Childs et al., 2016; Freeman et al., 2016; Muscott et al., 2008). The current quasi-experimental study is unique in that it eliminates many of the selection biases present in earlier research and includes a wide range of outcomes. The only other studies inclusive of similar methodology were conducted among a small sample within Minnesota and demonstrated neither academic nor behavioral effects of SW-PBIS (Ryoo et al., 2018) and focused only on discipline outcomes in a single year (Gage et al., 2019). Unlike many other preventive interventions, which may have the support of both efficacy and effectiveness research, SW-PBIS has been disseminated broadly and has population-level reach. Determining the extent to which it is impactful on student behavioral and academic outcomes throughout an entire state fills a gap in the extant knowledge about SW-PBIS and is relevant to schools throughout the United States and the world.

Over the six-year period that this study was conducted, SW-PBIS schools in the state saw improvements both on behavioral and academic indicators. Specifically, schools trained in SW-PBIS demonstrated improvements that were statistically significantly greater than those schools that were not trained in SW-PBIS. This finding was true both for the elementary and secondary schools across the range of targeted outcomes. In elementary schools, statistically significant improvements in suspensions and reading and mathematics proficiency were detected for schools implementing SW-PBIS for multiple years examined. The effects for suspensions in elementary schools were small, whereas the effect sizes for academic proficiency were medium to large for reading and ranged from small to large for math in different years. In secondary schools, these findings were isolated to two specific and early, within the study, school years (i.e., 2007-08 and

2008-09). Truancy was also improved by SW-PBIS in secondary schools. The effect sizes for suspensions in secondary schools were small but were medium for truancy and math and large for reading proficiency. SW-PBIS did not improve truancy rates in elementary schools.

Although the suspension findings were statistically significant, the notably small effect sizes on suspensions may be related to the whole-state decline in suspensions over the course of the study. Further, though we hypothesized behavioral outcomes to be proximal and academic outcomes to be distal, we found that behavioral and academics findings occurred simultaneously, and that there were larger effects for academics. This could be the result of the more objective and consistent nature of data collection for academic outcomes as compared to the behavioral outcomes or the relative greater room for growth on academic outcomes than behavioral outcomes; for example, academic proficiency rates were below 80% at the start of the study, allowing for over 20% improvement, whereas base rate suspensions were 11% and lower and were under 10% for truancy. Further, state trends in suspension rates also may have hindered the possible impacts SW-PBIS distinctly could make on this outcome specifically. The statistically significant results reported in the current study are consistent with those detected in RCTs of PBIS (Bradshaw et al., 2010; Horner et al., 2009), as well as a prior generalizability study associated with the Maryland RCT of SW-PBIS (Stuart et al., 2015), which indicated that trial results should generalize to the entire state; this prior generalizability study, coupled with the current study, provide further support for the potential broader impact of SW-PBIS on students within the state (Shadish et al., 2001).

We also considered the extent to which trained schools reached high fidelity implementation of SW-PBIS to contextualize the outcome analyses and to ensure that training could be equated with implementation. It is important to note that the trained schools in the

current study received high fidelity scores (i.e., over 80% in nearly all years across school types and measures; Horner et al., 2004). The differential averages on the IPI versus the SET likely stem from the fact that the IPI includes an assessment of late-stage implementation, including maintenance, and thus would take longer for high scores to emerge. Another important difference to note between the two measures is that the IPI is completed by the schools' coach, whereas the SET is completed by a trained external assessor, and thus may represent a more objective assessment of the schools' SW-PBIS implementation status. We included both of these measures in the analyses because of these measurement differences, thereby taking a conservative and inclusive approach to fidelity assessment. Regardless, high levels of intervention fidelity are often hard to achieve within the context of scale-up; other research suggests that preventive programs often suffer from poor implementation fidelity, particularly when implemented at scale (Gottfredson & Gottfredson, 2002; Rohrbach et al., 2006). Additional research focused only on SW-PBIS trained schools is needed to determine the extent to which the longitudinal effects vary by implementation fidelity over time, as prior research indicated that elementary and middle schools with a high PBIS fidelity had better attendance and academic outcomes (Pas & Bradshaw, 2012). Our interest for this paper was in contrasting trained and non-trained schools on impacts. Fidelity analyses would solely focus on intervention schools only, as comparison schools did not provide fidelity data, and the inherent added complexity of such fidelity analyses precluded us from conducting those here. Pas and colleagues (2019) examined the association between specific SW-PBIS fidelity cut points and student outcomes and reported that specific subscales within the SET measure correspond differently with behavioral and academic outcomes, and that simply examining overall fidelity or assuming consistent relationships between fidelity and the full range of student outcomes is not adequate.

A separate fidelity analysis for a state-wide scale up would be an important contribution to the field but should consider the complex, longitudinal nature of these data and examine the nuanced interplay between years of experience with SW-PBIS, the development of fidelity over time, and the way in which the emergence of fidelity coincides with a range of student outcomes.

Limitations and Future Directions

This study contributes to the knowledge base around the effectiveness of SW-PBIS but has some limitations to consider. In focusing specifically on the school-wide implementation of PBIS, we are not able to address the more advanced tiers. At the time of the data collection for this study, there was no statewide infrastructure for training in these tiers; large-scale training and the annual measurement of Tier 2 or 3 implementation did not begin until well after 2012. There is still a great need to assess implementation across all three tiers, which is the more recent and current focus of the state's current PBIS-related efforts. Maryland is one of a few states that was an early adopter of the SW-PBIS model, beginning to build the foundations for implementation in 1999 (see Barrett et al., 2008; Bradshaw et al., 2012; Bradshaw & Pas, 2011; Pas & Bradshaw, 2014). Taken together with the high levels of implementation achieved in this timeframe, it is possible that results in Maryland will not generalize to other states. In fact, the study conducted by Ryoo and colleagues (2018) did not find any behavioral or academic impacts of SW-PBIS. Additional replication research is needed to conclude whether such effects are generalizable beyond Maryland.

We merged the middle and high schools into a set of secondary schools, to optimize power and balance across the matched schools and reduce the number of statistical tests, but future analyses could explore the extent to which the effects differ for middle versus high schools. Research that explores differential effectiveness for middle schools as compared to high

schools would also further inform the field, as there has been limited research to date on the effectiveness of PBIS in high schools (e.g., Bradshaw, Debnam, et al., 2014). Similarly, we did not examine nesting at the district level, as the schools were nested within just 24 districts, the number of schools within districts varied considerably, and the current models included freed parameters exceeding the number of clusters. However, prior exploration of district-level factors and their association with schools seeking training in or adopting SW-PBIS yielded relatively few significant findings (i.e., the percent of schools trained in PBIS in the district and district size) and no such associations were found with fidelity scores (Bradshaw & Pas, 2011). This is a potential area to explore further in future analyses. In addition, moderation and mediation analyses are other areas for future research.

The use of propensity score methods is a strength, as PSM are a rigorous methodological approach to improve capacity for causal inference in the absence of randomization; however, biases (termed the propensity score matching paradox; King & Nielsen, 2016) can remain in the estimates resulting from PSMs. Some specific approaches are more vulnerable (e.g., matching) to this paradox than others. Propensity score weighting is not as vulnerable (King & Nielsen, 2016), which is why we employed weights instead of propensity score matching. To further promote bias reduction and elimination, we considered propensity score models that included covariates that were identified as good predictors in other SW-PBIS studies, so as not to result in a model suffering from model dependence and imbalance, both of which could affect bias. Finally, we identified the mean difference reductions with observational data to ensure improvements and balance were achieved. The weighted findings suggested only small differences (effect size of 0.20 or smaller) on all variables except suspensions rates in secondary schools during the 2004-05 school year. On the other hand, school-level factors such as buy-in

and organizational health, were not captured in this study, leaving plausible selection biases that we have not accounted for. This study represents an improvement over extant dissemination research in its inclusion of a comparison group and the weights do eliminate some selection bias; this still does not fully allow for causal inferences to be drawn.

It is unclear why SW-PBIS effects were statistically significant across years in elementary schools, but only during two specific years in secondary schools. It is possible that this relates to the varying levels of fidelity achieved in these two types of schools or that ceiling effects were being reached differentially in these school types. As hypothesized, we did see some early improvements in suspensions, but also in academic proficiency. It is possible that the improvements in all outcomes reached a point where further variability was limited after this time frame, and thus, power was limited to detect statistically significant differences between the two targeted groups. Specifically, there was a steady decline in suspensions across all years, for all schools. Similarly, the change scores indicated that all schools (but more notably, secondary schools) had a substantial increase in academic proficiency rates in 2007-08 followed by much less change generally after that point. The state had relatively recently adopted the state-wide achievement tests analyzed in this study period. It is possible that some of the state-wide improvements in academic proficiency overall in the state occurred as the result of the state adopting the new assessment format just a few years prior to the initial data point included in this study; as a result, schools experienced improvements in academic proficiency as they adjusted their curriculum to match the standards set by the new test and became more accustomed to this particular assessment by the 2007-08 school year. Despite this, significant improvements in behavior and academic achievement were found to be related to PBIS implementation. On the other hand, the findings still indicate differential improvements in academic proficiency,

favoring schools trained in SW-PBIS. SW-PBIS alone may not be able to continuously improve academic outcomes; it is likely that additional instructional interventions or practices would be needed to continue such growth and could explain the early and brief improvements demonstrated in secondary schools. Future research should explore this further.

On the other hand, the behavioral outcomes (i.e., suspensions and truancy) suggest room for further improvement. Further examination of the data at the student level is needed to determine whether suspensions are widespread across students (i.e., many students cause the still high number of events) or are among a relatively small subset of students who are not responding to the universal supports and need additional targeted or intensive interventions. For truancy, the rates demonstrate that a targeted population of students needs additional support to improve school attendance. Student-level data would also be needed to ascertain whether students who are truant also receive suspensions. Regardless, additional targeted and intensive interventions and practices are likely needed to decrease these rates further. For example, targeted engagement interventions, such as Check & Connect (Christenson et al., 2008), may be needed to engage students and further improve the suspension and truancy outcomes.

Conclusions and Implications

Given the wide-scale dissemination of SW-PBIS across U.S. schools, the findings of this study are particularly important and timely for educators and policymakers. There is also increased emphasis on school climate and use of proactive behavior management strategies in federal policies, like the Every Student Succeeds Act (ESSA, 2015). As a result, the current findings regarding the state-wide impact of SW-PBIS are particularly relevant, as they suggest that local education agencies and school districts should consider SW-PBIS as a research-based approach for improving a range of student behavioral and academic outcomes. The positive

effects observed in prior RCTs were largely replicated through this state-wide effectiveness study utilizing a quasi-experimental design. The effects appear to be particularly robust for elementary schools, as three out of the four targeted outcomes were consistently and statistically significant across multiple years. The findings in the secondary schools were less consistent across years, but still are promising. Although replication in other states would strengthen these conclusions, the consistency of these results with trials and non-experimental dissemination studies is compelling.

Efforts to articulate the benefit to costs ratio of PBIS are currently underway, both within the context of PBIS scale-up and under more controlled conditions of a RCT. For example, although the effect sizes for suspensions, as estimated based on prior RCTs of PBIS, might be interpreted as relatively small, the overall impact on reduced risk for high school completion suggests a relatively high cost savings for schools implementing the model with fidelity (see Swain-Bradway, Lindstrom Johnson, Bradshaw, & McIntosh, 2017). Similarly, methodologists have benchmarked the effect sizes for academic performance into typical expectations for growth in academic performance over time months of learning, which suggest that the effect sizes for academics observed in the current study are on par with, if not slightly larger than, some observed in other preventive interventions (e.g., Hill, Bloom, Black, & Lipsey, 2007). Finally, even the relatively modest effects of PBIS for an individual school are substantial when considering the combined impact across multiple student outcomes across the entire population of 26,000 schools in the U.S. that are trained in SW-PBIS.

ACKNOWLEDGEMENTS: The authors would like the thank the Maryland PBIS Management Team, which includes the Maryland State Department of Education, Sheppard Pratt Health System, and the 24 local school districts. We give special thanks to Philip Leaf, Katrina Debnam, Elizabeth Stuart, Joseph Kush, Kristina Kyles-Smith, Susan Barrett, and Jerry Bloom.

FUNDING: Support for this project comes from the Institute of Education Sciences (R305H150027).

References

Arden, S. V. Gruner Gandhi, A., Zumeta Edmonds, R., & Danielson, L. (2017). Toward more effective tiered systems: Lessons from national implementation efforts. *Exceptional Children, 83,* 269-280. doi: 10.1177/0014402917693565

Barrett, S. B., Bradshaw, C. P., & Lewis-Palmer, T. (2008). Maryland statewide PBIS initiative: Systems, evaluation, and next steps. *Journal of Positive Behavior Interventions, 10,* 105-114. doi: 10.1177/1098300707312541

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Quantitative Methods in Psychology, 107,* 238-246. doi: 10.1037/0033-2909.107.2.238

Bradshaw, C. P., Debnam, K., Lindstrom Johnson, S., Pas, E. T., Hershfeldt, P., Alexander, A., . . . Leaf, P. (2014). Maryland's evolving system of social, emotional, and behavioral interventions in the public schools: The Maryland Safe and Supportive Schools Project. *Adolescent Psychiatry, 4,* 194-206. doi: 10.2174/2210676604031409121631 20

Bradshaw, C. P., Debnam, K. J., Koth, C. W., & Leaf, P. J. (2009). Preliminary validation of the Implementation Phases Inventory for assessing fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions, 11*, 145-160. doi: 10.1177/1098300708319126

Bradshaw, C. P., Koth, C. W., Bevans, K. B., Ialongo, N. S., & Leaf, P. J. (2008). The impact of school-wide Positive Behavioral Interventions and Supports (PBIS) on the organizational health of elementary schools. *School Psychology Quarterly, 23,* 462-473. doi: 10.1177/1098300709334798

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering school climate through school-wide Positive Behavioral Interventions and Supports: Findings from a

group-randomized effectiveness trial. *Prevention Science, 10,* 100-115. doi: 10.1007/s11121-008-0114-9

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010). Examining the effects of schoolwide Positive Behavioral Interventions and Supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 12,* 133-148. doi: 10.1177/1098300709334798

Bradshaw, C. P., & Pas, E. T. (2011). A state-wide scale-up of Positive Behavioral Interventions and Supports (PBIS): A description of the development of systems of support and analysis of adoption and implementation. *School Psychology Review, 40*, 530-548.

Bradshaw, C. P., Pas, E. T., Barrett, S., Bloom, J., Hershfeldt, P., Alexander, A., . . . Leaf, P. (2012). A state-wide partnership to promote safe and supportive schools: The PBIS Maryland Initiative. *Administration and Policy in Mental Health and Mental Health Services Research, 39,* 225-237. doi: 10.1007/s10488-011-0384-6

Bradshaw, C. P., Pas, E. T., Goldweber, A., Rosenberg, M., & Leaf, P. J. (2012). Integrating School-wide Positive Behavioral Interventions and Supports with tier 2 coaching to student support teams: The PBISplus model. *Advances in School Mental Health Promotion, 5,* 177-193. doi: 10.1080/1754730X.2012.707429

Bradshaw, C. P., Reinke, W. M., Brown, L. D., Bevans, K. B., & Leaf, P. J. (2008). Implementation of school-wide Positive Behavioral Interventions and Supports (PBIS) in elementary schools: Observations from a randomized trial. *Education & Treatment of Children, 31,* 1-26. doi: 10.1353/etc.0.0025

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2012). Effects of School-Wide Positive Behavioral Interventions and Supports on child behavior problems. *Pediatrics, 130*, 1136-1145. doi: 10.1542/peds.2012-0243

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). New Bury Park, CA: Sage.

Childs, K. E., Kincaid, D., George, H., & Gage, N. A. (2016). The relationship between School-Wide Implementation of Positive Behavior Intervention and Supports and student discipline outcomes. *Journal of Positive Behavior Interventions, 18*, 89-99. doi: 10.1177/1098300715590398

Christenson, S. L., Thurlow, M. L., Sinclair, M. F., Lehr, C. A., Kaibel, C. M., Reschly, A. L., . . . Pohl, A. (2008). *Check & Connect: A comprehensive student engagement intervention manual*. Minneapolis, MN: University of Minnesota, Institute on Community Integration.

Cohen, J. (Ed.). (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum.

Every Student Succeeds Act of 2015, Pub. l. no. 114-95 § 114 stat. 1177 (2015).

Freeman, J., Simonsen, B., McCoach, B., Sugai, G., Lombardi, A., & Horner, R. H. (2016). Relationship between School-Wide Positive Behavior Interventions and Supports and academic, attendance, and behavior outcomes in high schools. *Journal of Positive Behavior Interventions, 18*, 41-51. doi: 10.1177/1098300715580992

Gage, N. A., Grasley-Boy, N., George, H. P., Childs, K., & Kincaid, D. (2019). A quasi-experimental design analysis of the effects of School-Wide Positive Behavior

Interventions and Supports on discipline in Florida. *Journal of Positive Behavior Interventions, 21,* 50-61. doi: 10.1177/1098300718768208

Gottfredson, G. D., & Gottfredson, D. C. (2002). Quality of school-based prevention programs: Results from a national survey. *Journal of Research in Crime and Delinquency, 39,* 3-35. doi: 10.1177/00224278020390010.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M.W. (2007). *Empirical Benchmarks for Interpreting Effect Sizes in Research.* MDRC Working Papers on Research Methodology. MDRC, New York, NY.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology, 2,* 259-278.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71,* 1161-1189.

Horner, R. H., Kincaid, D., Sugai, G., Lewis, T. J., Eber, L., Barrett, S., . . . Johnson, N. (2014). Scaling up School-Wide Positive Behavioral Interventions and Supports: Experiences of seven states with documented success. *Journal of Positive Behavior Interventions, 16,* 197-208. doi: 10.1177/1098300713503685

Horner, R. H., Sugai, G., Todd, A. W., & Lewis-Palmer, T. (2005). School-wide positive behavior support. In L. Bambara & L. Kern (Eds.), *Individualized supports for students with problem behaviors: Designing positive behavior plans* (pp. 359-390). New York: Guilford Press.

Horner, R. H., Sugai, G., & Anderson, C. M. (2010). Examining the evidence base for schoolwide positive behavior support. *Focus on Exceptional Children, 42,* 1-14.

Horner, R. H., Sugai, G., Kincaid, D., George, H., Lewis, T. J., Eber, L., . . . Algozzine, B. (2012). *What does it cost to implement School-Wide PBIS?* Retrieved from http://www.pbis.org/common/pbisresources/publications/20120802_WhatDoesItCostToImplementSWPBIS.pdf

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide Positive Behavior Support in elementary schools. *Journal of Positive Behavior Interventions, 11*, 133-144. doi: 10.1177/1098300709332067

Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The School-Wide Evaluation Tool (SET): A research instrument for assessing School-Wide Positive Behavior Support. *Journal of Positive Behavior Interventions, 6*, 3-12. doi: 10.1177/10983007040060010201

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi: 10.1080/10705519909540118

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. Retrieved from https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching

Kline, R. B. (2016). *Principles and Practices of Structural Equation Modeling* (4th ed.). New York, NY: The Guilford Press.

Little, T. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using

generalized boosted models. *Statistics in Medicine, 32*, 3388-3414. doi: 10.1002/sim.5753

McCaffrey, D. F., Ridgeway, G. & Morral, A.R. (2004). Propensity score estimation with boosted regressions for evaluating causal effects in observational studies. *Psychological Methods, 9,* 403-425. doi: 10.1037/1082-989X.9.4.403

Muscott, H., Mann, E., & LeBrun, M. R. (2008). Positive Behavioral Interventions and Supports in New Hampshire: Effects of large-scale implementation of schoolwide Positive Behavior Support on student discipline and academic achievement. *Journal of Positive Behavior Interventions, 10,* 190-205. doi: 10.1177/1098300708316258

Muthén, L. K., & Muthén, B. O. (2002-2018). *Mplus user's guide*. Retrieved from http://www.statmodel.com/ugexcerpts.shtml

O'Connell, M. E., Boat, T., & Warner, K. E. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, DC: The National Academies Press.

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes: State-wide scale-up of School-Wide Positive Behavior Intervention and Supports. *Journal of Behavioral Health Services and Research, 39,* 417-433. doi: 10.1007/s11414-012-9290-2

Pas, E. T., Johnson, S. R., Debnam, K. J., Hulleman, C. S., & Bradshaw, C. P. (2019). Examining the relative utility of PBIS implementation scores in relation to student outcomes. *Remedial and Special Education, 40*, 6-15. doi: 10.1177/074193251880519

Ridgeway, G., McCaffrey, D. F., Morral, A., Griffin, B. A., & Burgette, L. F. (2016). TWANG: Toolkit for Weighting and Analysis of Nonequivalent Groups. Retrieved from https://CRAN.R-project.org/package=twang

Rogers, E. M. (2002). Diffusion of preventive innovations. *Addictive Behaviors, 27*, 989-993.

Rohrbach, L. A., Grana, R., Sussman, S., & Valente, T. W. (2006). Type II Translation: Transporting prevention interventions from research to real-world settings. *Evaluation & The Health Professions, 29*, 302-333. doi: 10.1177/0163278706290408

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394. doi: 10.2307/2289440

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi: 10.1093/biomet/70.1.41

Ryoo, J. H., Hong, S., Bart, W., Shin, J., & Bradshaw, C.P. (2018). Investigating the effect of School-Wide Positive Behavioral Interventions and Supports on student learning and behavioral problems in elementary and middle schools. *Psychology in the Schools, 55*, 629-643.

Schoenwald, S. K., & Hoagwood, K. (2001). Effectiveness, transportability, and dissemination of interventions: What matters when? *Psychiatric Services, 52*, 1190-1197.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Houghton Mifflin.

Stuart, E., Bradshaw, C. P., & Leaf, P. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science, 16*, 475-485. doi: 10.1007/s11121-014-0513-z

Sugai, G., & Horner, R. H. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child & Family Behavior Therapy, 24*, 23-50. doi: 10.1300/J019v24n01_03

Sugai, G., & Horner, R. H. (2006). A promising approach for expanding and sustaining School-Wide Positive Behavior Support. *School Psychology Review, 35*, 245-259.

Sugai, G., Horner, R. H., & Gresham, F. M. (2002). Behaviorally effective school environments. In M. R. Shinn, H. M. Walker & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches.* (pp. 315-350). Bethesda, MD: National Association of School Psychologists.

Sugai, G., Horner, R. H., & McIntosh, K. (2016). *Multi-tiered systems of behavior support: Implementing PBIS at scales of social significance*. Retrieved from https://www.pbis.org/Common/Cms/files/pbisresources/MTSS-B-Implement%202016%2005-16.pdf

Sugai, G., Lewis-Palmer, T., Todd, A. W., & Horner, R. H. (2001). *School-wide Evaluation Tool (SET)*. Center for Positive Behavioral Supports, University of Oregon: Eugene, OR.

Swain-Bradway, J., Lindstrom Johnson, S., Bradshaw, C., & McIntosh, K. (2017, November). *What are the economic costs of implementing SWPIBS in comparison to the benefits from reducing suspensions?* PBIS Technical Assistance Center. University of Oregon, Eugene, OR.

Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of School-wide Positive Behavioral Interventions and Supports (SWPBIS) on bullying and peer rejection: A randomized controlled effectiveness trial. *Archives of Child and Adolescent Medicine, 116*, 149-156. doi: 10.1001/archpediatrics.2011.755

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-707. doi: 10.1146/annurev.soc.25.1.659

Table 1.
Standardized mean differences before and after propensity scores for both elementary and secondary schools in the first year of the study

	Elementary						**Secondary**					
	Unweighted means			Weighted means			Unweighted means			Weighted means		
	Int.	Control	*ES*	Int.	Control	*ES*	Int.	Control	*ES*	Int.	Control	*ES*
Enrollment	476.98	456.25	0.14	476.98	473.18	0.03	949.70	1118.67	**-0.41**	949.70	1029.31	-0.19
% Receiving special education	12.52	11.53	**0.21**	12.52	12.40	0.03	11.40	11.59	-0.05	11.40	11.31	0.02
% Receiving free and reduced meals	45.20	37.78	**0.31**	45.20	43.25	0.08	29.37	27.61	0.09	29.37	28.86	0.03
% Mobility	27.41	23.77	**0.22**	27.41	26.31	0.07	23.12	21.89	0.07	23.12	20.52	0.14
Student-Teacher Ratio	19.95	19.69	0.06	19.95	19.80	0.03	19.20	21.24	**-0.61**	19.20	19.78	-0.18
% American Indian	0.50	0.41	0.16	0.50	0.48	0.05	0.38	0.32	0.19	0.38	0.35	0.10
% Asian	4.26	5.55	**-0.25**	4.26	4.60	-0.07	4.34	5.49	**-0.24**	4.34	5.01	-0.14
% Hispanic	8.19	10.08	-0.17	8.19	8.76	-0.05	6.99	6.82	0.02	6.99	6.83	0.02
% African American	43.43	37.32	0.20	43.43	42.29	0.04	33.15	35.77	-0.10	33.15	34.24	-0.04
% White	43.62	46.27	-0.09	43.62	43.70	0.00	54.29	49.95	0.15	54.29	53.00	0.04
Suspension	7.16	3.29	**0.43**	7.16	5.86	0.14	26.71	18.85	**0.42**	26.71	22.72	**0.21**
Truancy	8.17	6.57	**0.25**	8.17	7.70	0.07	13.83	16.52	**-0.35**	13.83	15.26	-0.19
Reading	73.33	77.83	**-0.35**	73.33	74.51	-0.09	66.37	54.87	**0.54**	66.37	63.19	0.15
Math	69.23	74.51	**-0.33**	69.23	70.14	-0.06	59.46	51.73	**0.39**	59.46	58.21	0.06

Note. Int. = trained in PBIS. Suspension, truancy, and reading and math proficiency data were included from 2004-05 and all others were 2006-07 data. Standardized mean differences were measured and reported as an indicator of effect size (*ES*); *ES*s of less than 0.20 are considered ‘small’, 0.40 are considered ‘moderate’, and 0.60 are considered ‘large’; **bold** notes ‘moderate’ and ‘large’ ESs.

Table 2.
Implementation and training data across all study years

Year	Cumulative % Trained	IPI Total Score			SET Score		
All Schools		*M*	*SD*	Range	*M*	*SD*	Range
2005-06	19.6	79.3	17.2	32-100	84.2	21.2	13-100
2006-07	27.0	83.8	14.9	27-100	92.6	8.5	51-100
2007-08	37.4	81.6	16.9	18-100	94.9	6.4	68-100
2008-09	47.1	85.5	14.7	30-100	94.7	6.7	57-100
2009-10	55.6	86.9	14.0	31-100	95.7	5.7	53-100
2010-11	64.8	87.5	13.8	19-100	94.2	7.7	52-100
2011-12	73.2	87.7	14.5	22-100	94.0	8.9	32-100
2012-13	78.9	87.0	15.3	3-100	94.7	7.9	23-100
Elementary schools							
2005-06	18.1	83.6	15.9	32-100	81.9	24.7	13-100
2006-07	25.4	86.7	13.4	51-100	93.3	8.1	51-100
2007-08	34.9	84.2	15.1	22-100	95.9	5.7	68-100
2008-09	44.0	88.0	12.8	30-100	96.0	5.1	70-100
2009-10	52.8	88.7	13.0	31-100	96.3	4.9	71-100
2010-11	61.7	89.3	12.7	30-100	95.0	7.3	55-100
2011-12	69.6	90.2	11.6	28-100	95.1	7.8	32-100
2012-13	74.7	89.3	13.2	3-100	96.3	4.8	65-100
Secondary schools							
2005-06	22.2	74.1	17.6	43-100	88.1	11.9	35-100
2006-07	30.0	78.7	16.3	27-100	91.2	9.1	61-100
2007-08	42.1	77.3	18.9	18-100	93.0	7.2	70-100
2008-09	53.2	81.5	16.5	32-100	92.4	8.5	57-100
2009-10	60.9	83.6	15.1	36-100	94.4	6.9	53-100
2010-11	70.7	84.5	15.0	19-100	92.8	8.4	52-100
2011-12	80.1	83.6	17.8	22-100	91.7	10.4	50-100
2012-13	86.9	83.0	17.8	13-100	91.7	10.9	23-100

Note. IPI = Implementation Phases Inventory measure (total score is the average of the four scale scores). SET = School-wide Evaluation Tool. Both data points reflect measures collected in the spring of the given school year.

Table 3.
Baseline outcome rates and annual difference scores across all study years

	Suspensions		Truancy		Math		Reading	
All schools	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2005-06	10.49	14.51	9.28	8.52	72.82	17.45	74.96	16.11
Difference scores								
2006-07	0.54	6.29	-0.26	3.30	2.28	6.56	1.47	5.91
2007-08	-0.92	6.73	-0.33	3.17	5.83	11.69	7.87	10.78
2008-09	-1.11	7.05	-0.24	3.18	1.51	4.92	1.62	4.05
2009-10	-0.52	6.28	0.24	3.58	0.93	5.26	-0.27	4.41
2010-11	0.12	6.39	1.00	3.41	-0.14	5.81	0.62	4.73
2011-12	-0.68	5.77	-0.52	3.23	1.32	4.56	-0.28	3.80
Elementary schools								
2005-06	4.54	6.33	6.55	4.81	76.59	15.15	78.02	13.21
Difference scores								
2006-07	0.18	3.82	-0.26	2.71	3.10	6.20	2.19	5.31
2007-08	-0.41	4.33	-0.32	2.76	2.98	5.67	5.11	5.19
2008-09	-0.43	4.36	0.05	2.85	1.00	5.21	1.00	4.17
2009-10	-0.05	3.52	0.51	2.93	1.58	5.52	-0.12	4.64
2010-11	0.15	3.52	1.33	3.11	-0.50	5.77	0.77	4.78
2011-12	-0.01	3.26	-0.79	2.37	1.35	4.58	-0.02	3.67
Secondary schools								
2005-06	22.83	18.35	14.89	11.31	65.17	19.31	68.76	19.45
Difference scores								
2006-07	1.31	9.54	-0.29	4.29	0.53	6.87	0.04	6.71
2007-08	-2.00	9.95	-0.31	3.86	11.58	17.37	13.33	15.87
2008-09	2.48	10.52	-0.82	3.71	2.55	4.08	2.89	3.49
2009-10	-1.49	9.69	-0.31	4.60	-0.36	4.42	-0.49	3.71
2010-11	0.06	9.95	0.32	3.88	0.67	5.76	0.37	4.60
2011-12	-2.07	8.69	0.01	4.44	1.28	4.50	-0.77	3.92

Note. Averages and standard deviations for the rates of suspensions, truancy, and proficiency on the Maryland School Assessments (MSA)/High School Assessment (HSA) in math/algebra and reading/English language arts are reported for the 2005-06 school year. Average difference scores and the standard deviations of these difference scores are provided for each subsequent year. Difference scores were calculated by subtracting the given year's rate from the prior year's rate on each outcome.

Table 4.
SW-PBIS effects in elementary and secondary schools from 2006-07 to 2011-12.

Year	Elementary								Secondary							
	Suspension		Truancy		Reading		Math		Suspension		Truancy		Reading		Math	
	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*
SW-PBIS effects																
2006-07	-0.40	0.27	-0.31	0.14	**0.98**	**0.02**	**2.38**	**0.00**	0.48	0.61	-0.44	0.18	0.23	0.62	0.75	0.12
2007-08	-0.36	0.36	-0.21	0.26	**3.65**	**0.00**	**1.30**	**0.00**	**-2.25**	**0.01**	**-1.33**	**0.00**	**6.10**	**0.00**	**5.37**	**0.00**
2008-09	-0.40	0.15	-0.09	0.64	0.39	0.23	0.08	0.81	0.64	0.58	-0.46	0.12	**1.67**	**0.00**	**1.11**	**0.01**
2009-10	**-0.53**	**0.02**	-0.41	0.06	0.62	0.09	**1.38**	**0.00**	0.11	0.90	-0.23	0.51	-0.02	0.97	-0.19	0.70
2010-11	**-0.74**	**0.03**	0.06	0.80	**1.47**	**0.00**	0.92	0.13	-0.07	0.93	-0.14	0.70	0.38	0.43	0.85	0.15
2011-12	-0.48	0.09	0.00	0.99	0.04	0.91	**1.07**	**0.00**	0.08	0.95	0.65	0.29	1.26	0.06	0.27	0.74
Prior year's data																
2007-08	-0.28	0.06	**-0.39**	**0.00**	**-0.21**	**0.01**	-0.07	0.29	**-0.33**	**0.00**	-0.09	0.58	**-0.69**	**0.00**	**-0.88**	**0.00**
2008-09	**-0.35**	**0.00**	**-0.26**	**0.01**	**-0.20**	**0.00**	**-0.12**	**0.03**	**-0.25**	**0.00**	**-0.38**	**0.00**	-0.03	0.12	0.01	0.62
2009-10	**-0.29**	**0.00**	**-0.27**	**0.01**	**-0.51**	**0.00**	**-0.36**	**0.00**	**-0.34**	**0.00**	**-0.60**	**0.00**	**-0.24**	**0.01**	0.01	0.93
2010-11	**-0.38**	**0.00**	**-0.49**	**0.00**	**-0.35**	**0.00**	**-0.23**	**0.00**	**-0.32**	**0.01**	**-0.30**	**0.02**	**-0.52**	**0.00**	**-0.32**	**0.00**
2011-12	**-0.30**	**0.00**	**-0.38**	**0.00**	-0.05	0.25	**-0.15**	**0.00**	**-0.62**	**0.00**	**-0.47**	**0.00**	**-0.32**	**0.01**	-0.13	0.18
Covariates																
Enroll	0.00	0.56	0.00	0.61	0.00	0.93	-0.00	0.30	0.00	0.29	0.00	0.13	0.00	0.33	0.00	0.72
FARMs	0.00	0.86	0.00	0.54	0.00	0.56	0.01	0.12	0.04	0.38	-0.01	0.45	**0.03**	**0.01**	**0.03**	**0.02**
Mobility	0.01	0.65	-0.01	0.16	0.00	0.70	0.01	0.19	-0.01	0.45	-0.00	0.12	0.00	0.35	0.00	0.35
% White	-0.00	0.50	-0.00	0.74	0.00	0.95	0.01	0.30	0.03	0.33	-0.00	0.67	**0.02**	**0.01**	**0.02**	**0.00**
Truancy	-0.01	0.91	NA	NA	0.03	0.50	**0.15**	**0.01**	0.01	0.75	NA	NA	0.01	0.49	0.00	0.99
Read	0.02	0.34	0.00	0.71	NA	NA	NA	NA	-0.04	0.09	-0.01	0.13	NA	NA	NA	NA
Math	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Susp.	NA	NA	0.03	0.35	0.06	0.05	0.03	0.54	NA	NA	0.01	0.40	**-0.03**	**0.01**	-0.01	0.40

Note. All intervention effects are reported, controlling for the listed covariates. Significant findings are bolded.

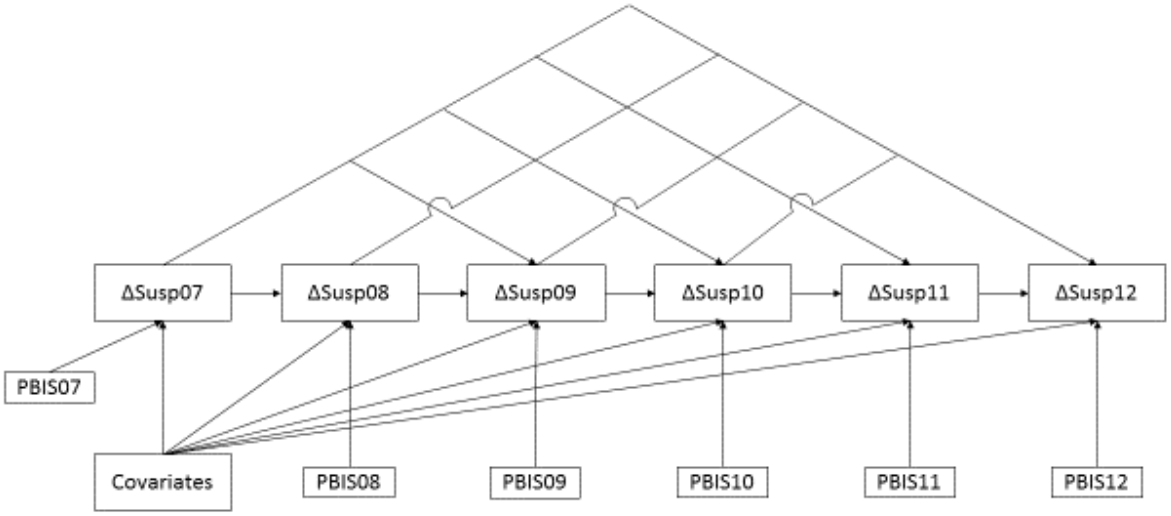
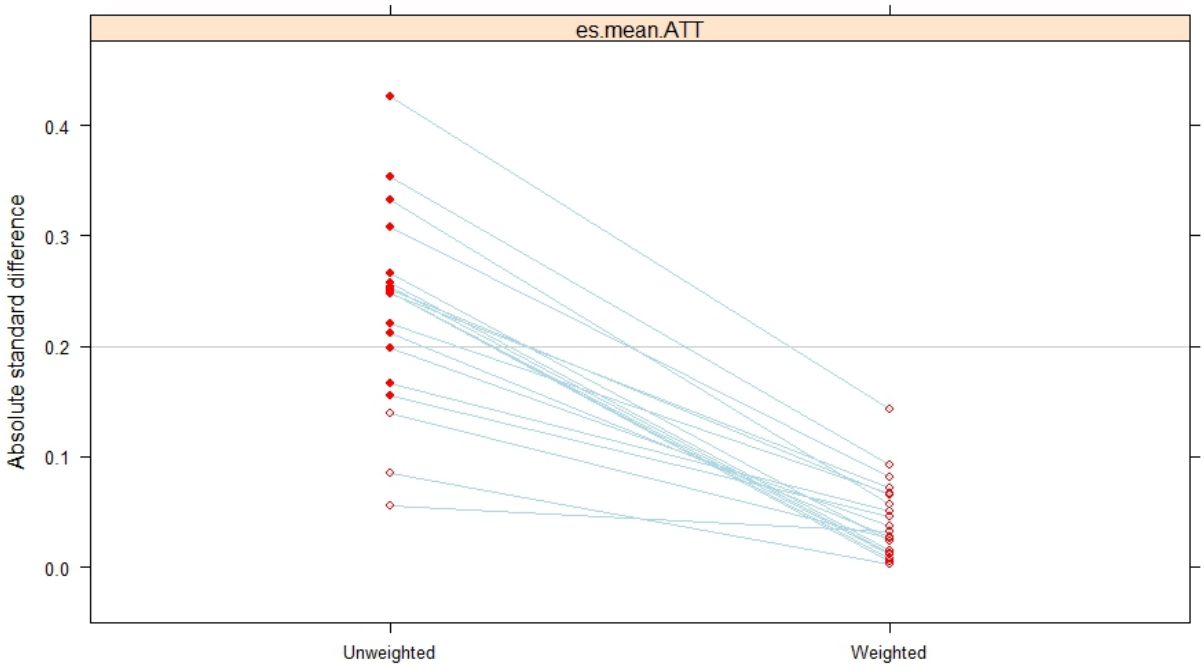
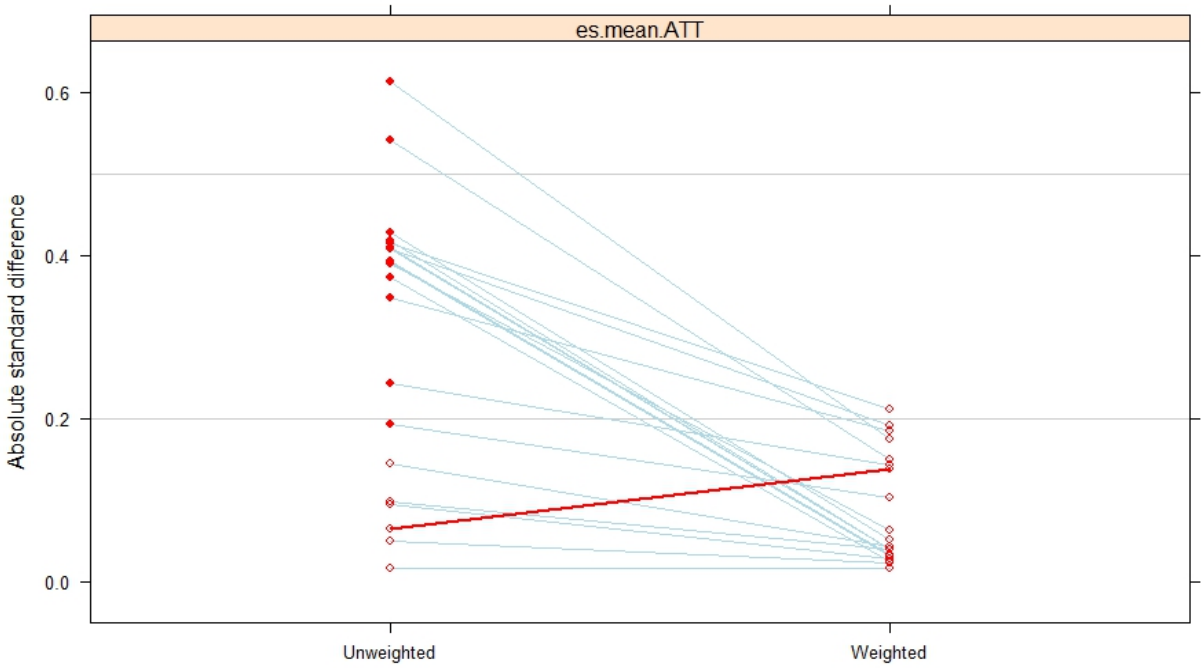


Figure 1. Panel model with autoregressive structure depicting the suspension outcome across 6 years where Δ indicates the weighted difference score, i.e., $\Delta\text{Susp07} = w_{i1} \cdot$ (Susp07-Susp06)$\Delta\text{Susp07} = w_{i1} \cdot$ (Susp07-Susp06), where $w_{i1}w_{i1}$ is the propensity score weight for *i*-th subject at time 1.



(a) Elementary schools



(b) Secondary schools

Figure 2. Reducing the mean differences for elementary schools (a) and secondary schools (b) using propensity score weights