Evaluating the Role of Classroom Behavior Management in Promoting Equitable Discipline Outcomes

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Alexandria C. Robers

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Dr. Faith G. Miller

July 2023

Acknowledgements

I want to first acknowledge my faith in God as a consistent source of strength throughout this process. I could not be where I am today without God and the most amazing mother in the world, Debbie Turner. Their love and wisdom have guided me through the most tumultuous moments in my life (this dissertation process included!). Ma, as your one and only daughter, I love you so much. I also want to acknowledge my partner in this life and husband, Christopher Robers. You know better than anyone else the long nights, tears, and missed together time that occurred because of this dissertation. I am incredibly blessed to have you as my biggest supporter throughout this journey and I love you for sticking by my side.

Next, I want to highlight one of the most cherished friendships I have ever been fortunate enough to experience: Intel. Dr. Jenna McGinnis and Sydney Pauling, you two mean the world to me. I appreciate y'all for keeping me in the group chat despite my delayed response time due to dissertating. The confidence you both had in me that I would get this done sustained me more than you could know. I love you both. The broader peer support I also received from best cohort ever (BCE) is something that I will never take for granted. I am working toward becoming the best school psychologist I can be because of BCE, thank you!

There is no way I would have been able to work full-time on internship and work on this milestone without the protected writing sessions with Marie Tanaka. I also want to acknowledge Mahasweta Bose and Greta Kos who took the time to learn how to code articles that were part of my study 1 results. I also need to give a special thank you to Drs. Mohammed Abulela and Rik Lamm who provided invaluable consultation and expertise regarding the statistical analyses for both studies.

Finally, I want to thank my committee for taking the time to review and provide feedback to my dissertation: Drs. Kim Gibbons, Annie Hansen-Burke, and Amanda Sullivan. I especially want to thank my advisor, Dr. Faith Miller, who has always supported my work and made me feel like my perspective belongs in our field. Faith, I am forever appreciative of your commitment to ensuring high-quality work is conducted. This dissertation reflects that commitment and your thoroughness. I am all the better for it, so thank you again.

Abstract

A common consequence for students who engage in disruptive behaviors at school is their removal from the instructional environment. Depending on the frequency and/or intensity of the disruptive behavior, this removal occurs from classrooms (e.g., office discipline referrals), school buildings (e.g., suspensions), or districts (e.g., expulsions). Regardless of the type of exclusionary discipline used, it is well-established that students of color, specifically Black males, are the most likely to be excluded from school due to their behavior. This finding persists even when these students behave similarly to White students. This is likely because, in addition to classroom teachers and other school professionals being underprepared to effectively manage these disruptive behaviors, they are also unable to do so in a culturally humble manner. While there exists various interventions and frameworks to help address the lack of knowledge and training school staff have on behavior management, it remains unknown how well these approaches prevent the occurrence of exclusionary discipline, especially in mitigating persistent racial discipline disparities. Thus, the purpose of this two-study dissertation is to further investigate the relation between implementation of classroom behavior management interventions and the promotion of equitable discipline outcomes. In study 1, a systematic review and quantitative synthesis were conducted to evaluate equitable discipline outcomes, as measured by the reduction in use of exclusionary discipline, after implementation of a specific classroom behavior management intervention called Classwide Function-related Intervention Teams (CW-FIT). In study 2, confirmatory factor analysis and multiple linear regression were conducted to evaluate equitable discipline outcomes, as measured by improvement of racial discipline disparities across specific racial and ethnic groups, after implementation of the positive behavioral interventions and support (PBIS) framework. Results from both of these studies and implications for practice and research are discussed.

Keywords: equitable school discipline, CW-FIT, PBIS, classroom behavior management

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

When students engage in disruptive behavior that interferes with classroom instruction, teachers tend to have difficulty effectively managing this behavior. A common classroom behavior management approach used by teachers is the removal of students from the classroom environment due to their disruptive behavior (i.e., exclusionary discipline). Exclusionary discipline practices can range in severity, from office disciplinary referrals (ODRs) to suspensions and expulsions; however, all result in loss of opportunity to learn and participate in the classroom environment. The use of these practices is disproportionate, with systematic reviews consistently indicating that students from certain racial and ethnic backgrounds experience exclusionary discipline more frequently than their White peers (e.g., Gopalan & Nelson, 2019; Young et al., 2018). This phenomenon, known as racial discipline disparities or the discipline gap, illustrates the need for solutions to address this issue. Although multifaceted, growing evidence supports a link between class-wide practices and the equitable use of exclusionary discipline practices (Bottiani et al., 2018). For example, teachers receiving classroom-based coaching (Bradshaw et al., 2018; Gion et al., 2020), implementing novel classroom behavior management interventions (Cook et al., 2018), and incorporating restorative practices in their classrooms (Gregory et al., 2016) all resulted in reduced disproportionality in exclusionary discipline, specifically reducing the odds, or risk ratio, of students of color receiving ODRs. Despite these promising findings, additional research is needed to evaluate the connection between classroom behavior management strategies and the (in)equitable use of exclusionary discipline practices in schools. To this end, the aims of this two-study dissertation were to: (1) evaluate the evidence supporting the use of a promising classroom behavior management intervention, the Classwide Function-related Intervention Teams (CW-FIT), in improving class-wide behavior and disparities in exclusionary discipline, and (2) examine the association between positive behavioral interventions and supports (PBIS) implemented in classrooms and improved equitable outcomes in school exclusionary discipline via ODRs.

Racial Discipline Disparities and Students of Color

For nearly half a century, researchers identified racial discipline disparities in the use of exclusionary discipline practices (Children's Defense Fund, 1975; Fenning & Rose, 2007; Gregory et al., 2011; Pane et al., 2014; Skiba et al., 2002; Skiba et al., 2011; Wallace et al., 2008) along with concomitant life-long negative outcomes associated with these practices, including school dropout, incarceration, and isolation from prosocial peers (Dupper, 2010; McIntosh et al., 2008; Robison et al., 2017). When youth disproportionately experience exclusionary discipline practices, this is referred to as the discipline gap (Bradshaw et al., 2010; Fenning & Rose, 2017; Sullivan et al., 2013). Although there are no significant differences in how youth from different racial and ethnic backgrounds behave (Skiba et al., 2011), compared to their White peers, these students are still more likely to experience exclusionary discipline practices in schools. In particular, students who are Black (Bradshaw et al., 2010) and American Indian/Native Alaskan (Vincent et al., 2012) consistently are the most likely to experience severe forms of exclusionary discipline like suspensions and expulsions. Conversely, Asian/Pacific Islander (Cruz & Rodl, 2018) students are the least likely to experience this form of exclusionary discipline. Outcomes for Latine students are inconsistent, with some evidence suggesting that these students are more likely to experience suspensions and expulsions (Skiba et al., 2011) whereas other evidence suggests that they are less likely (Morgan & Wright, 2018).

Racial Discipline Disparities and Office Discipline Referrals (ODRs)

In addition to suspensions and expulsions, other research studies evaluating racial disparities in school discipline practices, focus on a different type of exclusionary discipline outcome (i.e., ODRs). Classroom teachers often rely on ODRs to manage disruptive student behavior, but in addition to removing students from the classroom environment, ODRs limit student access to learning opportunities (Brown-Browner, 2019; Skiba et al., 1997; Tidwell et al., 2003). In many contexts, a student receiving an ODR serves as the student's initial experience with exclusionary discipline (Morrison et al., 2001). Over time, as students experience more severe forms of exclusionary discipline (i.e., suspensions) then they are more likely to experience life-long negative outcomes such as school dropout (Chu & Ready, 2018; Hirschfield, 2009; Lee et al., 2011). Consequently, addressing the use of ODRs serves as an entry point in potentially preventing the considerable negative outcomes associated with students who experience continued exclusion from school due to behavior. One approach that routinely focuses on the use of ODRs as a primary outcome of interest is PBIS.

Racial Discipline Disparities and Positive Behavioral Interventions and Supports (PBIS)

PBIS is a common implementation framework for school professionals to use to support student behavior. Key principles of PBIS include integrated data, systems, and practices that endeavor to promote prosocial behavior and prevent problem behavior. Several recent reviews show the effects of Tier 1 PBIS in improving both school and student outcomes in high school (Estrapala et al., 2020) and alternative education settings (Grasley-Boy et al., 2020). Implementation of PBIS is linked to improved school outcomes such as, school climate (Bradshaw et al., 2009) as well as student outcomes such as increased prosocial behavior (Bradshaw et al., 2010), decreased conflict amongst peers (Waasdorp et al., 2012), and a reduced number of ODRs delivered by school personnel (Bradshaw et al., 2012; Flannery et al., 2014; Molloy et al., 2013; Noltmeyer et al., 2019). Despite the positive outcomes associated with PBIS implementation in terms of reducing the number of ODRs across different school contexts, inequities in the use of ODRs remain (McIntosh et al., 2018; Vincent & Tobin, 2010).

Culturally Responsive PBIS

Due to the inconsistent evidence on the effects of PBIS in mitigating racial discipline disparities, there is a need for practices within PBIS systems to be culturally responsive to the unique and various needs of students (Sugai et al., 2012; Bastable et al., 2021). This is especially true as the demographic make-up of teachers and administrators who are working at schools does not reflect the demographic make-up of students who are attending those schools. There is a lot of theoretical support for the connection between predominantly White teachers disciplining predominantly non-White children in sustaining White supremacy, engaging in White saviorism, and committing to color-evasive racism (Bryan, 2017; Sondel et al., 2019). In addition, there is growing empirical support highlighting how racial disparities in the use of exclusionary discipline practices in schools relate to the disengagement of children in school, their eventual exit from the educational system, and an increased likelihood that they will enter into the juvenile justice system (Skiba et al., 2014). All of this further emphasizes the importance for those implementing frameworks like PBIS to adopt a culturally responsive and humble approach. That is, a color-evasive approach to behavioral support will perpetuate racial discipline disparities. Instead, school professionals must intentionally ensure their PBIS system is culturally responsive to the needs of their students with the goal to produce equitable outcomes and disrupt racial discipline disparities.

Vulnerable Decision Points (VDP) Model

Recently, a model known as the Vulnerable Decision Points (VDP) model was developed to further understand the phenomenon of racial discipline disparities within PBIS systems (McIntosh et al., 2014). The VDP model describes the conditions under which racial bias is likely to occur in the context of school discipline. Vulnerable decision points are specific situations or circumstances in which there is an increased likelihood that data within a PBIS system will be disproportionate. Specifically, discipline decisions made by teachers in the classroom may serve as a *vulnerable decision point* when it comes to the inequitable use of exclusionary discipline practices in schools (Smolkowski et al., 2016). That is, Smolkowski and colleagues (2016) note that the classroom context, especially during periods of time in which there is a strong academic focus, may be particularly linked to disproportionate use of ODRs. Therefore, addressing the use of ODRs in classrooms appears to be an important potential lever for change in mitigating racial discipline disparities. Consequently, there is a need for research to focus on evaluating the extent to which classroom practices improve racial discipline disparities that specifically exist in the use of ODRs in schools.

Racial Discipline Disparities and Classroom Behavior Management

Although *school-wide* approaches like setting behavior expectations and outlining discipline policies are effective, there is evidence that in order to have fewer biased discipline decisions in schools, the focus needs to be on implementing behavior management techniques in *classrooms* (Childs et al., 2016; Gregory et al., 2016; Mitchell & Bradshaw, 2013; Pane et al., 2015). As an alternative to removing students from the classroom environment, classroom behavior management strategies offer a promising approach for equipping teachers with strategies to promote more inclusive and equitable practices. However, because the outcomes of classroom behavior management strategies are often focused on improving specific student behaviors (e.g., time on task, work completion, disruptive behavior), it is unclear to what extent these strategies might improve equitable discipline outcomes in schools.

Rationale

The goal of this dissertation was to evaluate the interplay between classroom behavior management strategies and the (in)equitable use of exclusionary discipline practices, particularly as it relates to ODRs. The current state of the literature prioritizes evaluating classroom behavior management strategies and their effects on specific student-level behavioral outcomes (often measured by systematic direct observation) as well as investigating school-wide approaches in improving equitable discipline outcomes. However, investigating both of these approaches alone is insufficient. Recent evidence indicates that teachers' use of classroom behavior management strategies influences their decisions surrounding when to use ODRs which effects the discipline gap (Cook et al., 2018; Gregory et al., 2016; Gregory & Roberts, 2017). Thus, investigations that integrate how classroom behavior management strategies relate to the use of exclusionary discipline broadly, and ODRs specifically, are needed.

To this end, the purpose of the first study in this dissertation was to systematically review the literature of a packaged collection of classroom behavior management strategies, which included evaluating the extent to which outcome measures were reported to examine changes in the disproportionate use of exclusionary discipline practices. To accomplish the aims of this study, a systematic review and quantitative synthesis were conducted to summarize the effects of Tier 1 components of a classroom behavior management intervention, Classwide Function-Related Intervention Teams (CW-FIT), on the class-wide behaviors of students. An additional aim of the first study was to examine the extent to which outcome measures related to disproportionality in exclusionary discipline in schools were reported. The methods and approach in the first study were used to answer the following research questions:

Study 1:

1. To what extent does implementation of the CW-FIT, at Tier 1, relate to improved class-wide student behavioral outcomes?

2. To what extent does implementation of Tier 1 components of the CW-FIT relate to reduced use of exclusionary disciplinary practices delivered by teachers in the classroom?

The purpose of the second study in this dissertation was to determine if scores on a school-wide PBIS implementation fidelity measure, the Benchmarks of Quality Revised (BoQ-R), were associated with racial discipline disparities in ODRs. To accomplish the aims of this study, researchers utilized confirmatory factor analysis (CFA) to evaluate psychometric evidence for the BoQ-R. In addition, researchers conducted multiple linear regression to examine the association between scores from the BoQ-R and racial discipline disparities in ODRs. In particular, the BoQ-R measure includes a *Classroom Systems* critical element to evaluate the implementation of PBIS practices in classrooms. Scores from this critical element specifically permitted analysis of the association between practices that are aligned with recommended classroom behavior management strategies and racial disparities in the use of exclusionary discipline practices in schools. The methods and approach in the second study were used to answer the following research questions:

Study 2:

1. Is there evidence for structural validity of the BoQ-R in a national sample?
2. To what extent does implementation fidelity of PBIS (as measured by the total score) from the BoQ-R as well as implementation fidelity of PBIS components (as measured by certain critical elements) from the BoQ-R (e.g., *Classroom Systems*), relate to a decrease in the average number of ODRs given to specific racial and ethnic groups within a sample of U.S. schools?
3. To what extent does implementation fidelity of PBIS (as measured by the total score) from the BoQ-R as well as implementation fidelity of PBIS components (as measured by certain critical elements) from the BoQ-R (e.g., *Classroom Systems*), relate to greater equitable school discipline outcomes (i.e., risk ratio) across specific racial and ethnic groups within a sample of U.S. schools?

Researchers hypothesized that findings from this dissertation would help to identify additional components that relate to closing the discipline gap in schools. Specifically, the two studies aimed to link the influence of classroom behavior management strategies in improving class-wide student behavior, reducing the occurrence of exclusionary discipline practices, and closing the discipline gap. That is, by better equipping teachers with skills to support student behavior in the classroom, inequitable discipline practices might diminish.

Summary

Despite implementation of school-wide behavioral interventions and supports, students from certain racial and ethnic backgrounds (i.e., American Indian/Alaskan Native and Black) are consistently at greater risk for experiencing exclusionary discipline practices that disrupt access to learning opportunities. These exclusionary discipline disparities persist across all levels of the K-12 education system (Goplan & Nelson, 2019). The persistence of the discipline gap in schools makes it clear the need for multi-faceted solutions to solve this problem. Yet, despite more evidence supporting investigation of practices at the classroom-level, the majority of the research in this area focuses on influences of school- and student-level factors. Thus, the purpose of this dissertation was to examine how classroom behavior management practices, designed to manage disruptive student behavior, might also be implicated in the (in)equitable use of exclusionary discipline practices in schools.

Chapter 2

Promoting Positive Behavioral Outcomes in the Classroom: A Systematic Review and Quantitative Synthesis of Tier 1 Components of the CW-FIT

Classroom behavior management, or "the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior" (Pianta et al., 2008, p. 44), can help teachers address student behavior that disrupts the classroom environment. Whether it is relying on specific strategies (e.g., praise, opportunities to respond), or implementing packaged intervention programs (e.g., the Good Behavior Game), there are many options available for teachers and other school staff to improve their classroom behavior management skills. The Classwide Function-Related Intervention Teams (CW-FIT) is one such intervention program that combines evidence-based behavior management practices into an intervention package for teachers to implement in their classrooms (Wills et al., 2009). A unique and promising aspect of the CW-FIT is that it can be implemented across multiple tiers of support beyond Tier 1, and a recent review of classroom management strategies indicated potential promise of the program when used with racially and ethnically minoritized students (Long et al., 2019). While there are a number of studies evaluating the effectiveness of the CW-FIT, these studies have yet to be synthesized in order to better understand the state of the literature. Currently, it is unclear to what extent results from these studies effectively improved student behavior, were methodologically rigorous, and considered the effects of the CW-FIT on the use of exclusionary discipline practices. The purpose of the present study was to perform both a systematic review of empirical studies implementing the CW-FIT and a quantitative synthesis of results from these studies to better understand the body of evidence for the CW-FIT, especially as it relates to the use of exclusionary discipline practices in the classroom.

Managing Challenging Behavior in the Classroom

When students engage in challenging behavior that disrupts the instructional environment, classroom teachers can either implement proactive or reactive approaches to manage such behavior. Proactive strategies are prevention and antecedent-focused and aim to set up the classroom environment to minimize the occurrence of challenging behavior and promote engagement (Kern & Clemens, 2007;

Simonsen et al., 2008). Examples of proactive strategies include clearly stating, teaching, and reinforcing classroom behavior expectations, providing opportunities to respond, and arranging the environment to promote appropriate student behavior (e.g., seating charts, labeled stations). In contrast to proactive strategies, reactive strategies occur after the occurrence of challenging behavior. Specific reactive strategies include reprimands, corporal punishment, and exclusionary discipline. These reactive strategies aim to serve as punishers with the intent to decrease the future occurrence of challenging behavior. A noteworthy limitation of punishers is that punishment does not teach students the appropriate behaviors they are expected to engage in (Maag, 2001; Scott et al., 2007; Shores et al., 1993). Yet, the overreliance of reactive strategies, especially exclusionary discipline practices, in classrooms is evident in the persistent disparities in the use of exclusionary discipline practices in schools.

Exclusionary Discipline Practices in Schools

Exclusionary discipline practices are actions that remove the student from the instructional environment, including suspensions, expulsions, or office discipline referrals (ODRs) in response to "problem" behavior (Mitchell & Bradshaw, 2013). Despite their frequent use, associations between exclusionary discipline practices and negative life-long student outcomes such as school dropout, interaction with the juvenile justice system, and social isolation from prosocial peers are well-documented (Dupper, 2010; McIntosh et al., 2008; Robison et al., 2017). As such, it is prudent to address minimizing the use of exclusionary discipline by examining practices in the classroom setting. To this end, within each teacher resides the power to decide whether or not to exclude a student from the classroom (Girvan et al., 2016; McIntosh et al., 2014; Pane et al., 2014; Smolkowski et al., 2016). Still, much of what we know about exclusionary discipline practices involves analysis of major disciplinary incidents, such as suspensions and expulsions (e.g., Algozzine & Algozzine, 2007; Anyon et al., 2016; Bottiani et al., 2017; Fisher & Hennessy, 2016; Gopalan & Nelson, 2019; Marcucci, 2020; Mitchell & Bradshaw, 2013; Pearman II et al., 2019; Smith et al., 2020) as opposed to more covert exclusionary disciplinary practices initiated in classrooms (i.e., ODRs; Mizel et al., 2016). Indeed, exclusionary discipline often starts in the classroom (e.g., ODRs, time-out procedures), and eventually expands in scope to be school-wide (e.g., suspensions and expulsions; Petras et al., 2011). Thus, evaluations of behavior management interventions designed specifically for the classroom setting are warranted.

Classroom Behavior Management Interventions

Classroom behavior management interventions serve as an alternative to exclusionary discipline practices delivered by teachers in the classroom. Instead of a reactive and negative response to challenging behavior, as characterized by exclusionary discipline practices, classroom behavior management interventions are proactive and responsive to the behavioral needs of students. Classroom behavior management interventions differ from the more general concept of classroom management. In the present study, classroom management was broadly defined to include any actions that create and maintain an atmosphere in the classroom that is conducive to learning (Barbetta et al., 2005). Therefore, classroom behavior management was conceptualized as a component of classroom management that involves specific skills and techniques teachers can use to address student behavior that disrupts the classroom environment. This conceptualization was supported by the well-documented finding in the literature that pre-service teachers are underprepared to effectively manage classroom behavior that interferes with instruction (Baker, 2005; Eisenman et al., 2015; Freeman et al., 2014). Even when a school-wide approach is taken to address student misbehavior, teachers struggled to implement behavioral interventions and supports in the classroom (Fallon et al., 2014).

Fortunately, there exists numerous classroom behavior management interventions shown to effectively improve student behavior (Barbetta et al., 2005; Dougherty & Dougherty, 1977; Noell et al., 2002; Tillery et al., 2010). So much so, researchers investigated the utility of combining these individually effective strategies and techniques (Chafouleas et al., 2012; De Pry & Sugai, 2002; Reinke et al., 2008; Walker et al., 1981; Wills et al., 2009). The CW-FIT is an example of one such intervention package that combines multiple evidence-based classroom behavior management techniques into a comprehensive approach for teachers to manage student behavior.

Classwide Function-Related Intervention Teams (CW-FIT)

The CW-FIT is a multi-level group contingency intervention designed to address common functions of problem behavior in the classroom (Wills et al., 2009). The CW-FIT is divided into three levels, consistent with a multi-tiered system of support framework. Tier 1 is a group contingency which involves earning group or individual rewards based on engaging in desired behavior related to the instruction and extinction components of the CW-FIT. The instruction component involves teaching students targeted skills (i.e., following directions, obtaining teacher's attention, and disregarding inappropriate behavior) and the extinction component involves teaching students how to ignore peer problem behavior to prevent social reinforcement of such behavior (Wills et al., 2009). Tier 2 of the CW-FIT is an intervention designed for students whose behaviors are not improving after exposure to the Tier 1. This Tier 2 intervention involves teaching students self-management strategies that include providing students with opportunities to request peer tutoring (Wills et al., 2009). Tier 3 of the CW-FIT involves conducting a functional behavior assessment (FBA) of student behavior for those who did not respond to the interventions provided in Tiers 1 and 2 of the CW-FIT (Wills et al., 2009).

While various classroom behavior management interventions exist, the CW-FIT is unique and offers potential advantages above and beyond other approaches. For example, most classroom behavior management interventions are implemented exclusively at Tier 1, but the CW-FIT program has the potential to address the behavioral needs of all students, including the ones who need greater support than what Tier 1 can provide. Thus, the CW-FIT program could effectively support teachers in managing classroom behavior of varying severity. Furthermore, the CW-FIT encompasses a multifaceted approach and has the built-in capability to address the function of student problem behavior at multiple levels (Wills et al., 2009). That is, the intervention components included within and across each tier can address multiple behavioral functions. Consequently, the CW-FIT appears to be a promising approach to comprehensively support positive behavior in the classroom.

Although there are a number of individual research studies evaluating the effectiveness of the CW-FIT, there has not been a comprehensive review of the quality of these studies or an evaluation of the aggregate effectiveness of Tier 1 implementation of the CW-FIT in improving student behavior. At this time, it is unknown to what extent the CW-FIT can be classified as an evidence-based practice. Additionally, the CW-FIT has been implemented in classrooms with racially and ethnically minoritized students (Long et al., 2019), which may permit the examination of differential effects across various populations. Finally, because teachers serve as gatekeepers in deciding whether or not students are excluded from the instructional environment due to behavior, it is important to know the extent to which implementation of programs, like the CW-FIT, potentially minimize the use of teacher-delivered exclusionary discipline practices in the classroom. Hence, we chose to focus on Tier 1 implementation as it may serve as the gateway to exclusionary discipline practices.

Current Study

The purpose of the present study was to quantify the effectiveness of implementing Tier 1 components of the CW-FIT in improving class-wide student behavior and to examine whether studies have examined impacts on exclusionary discipline practices. As such, a systematic review of the literature and a quantitative synthesis were conducted. To that end, the present study aimed to answer the following research questions:

1. To what extent does implementation of the CW-FIT, at Tier 1, relate to improved class-wide student behavioral outcomes?
2. To what extent does implementation of Tier 1 components of the CW-FIT relate to reduced use of exclusionary discipline practices delivered by teachers in the classroom?

Because the CW-FIT is a collection of evidence-based classroom behavior management practices, it was expected that the class-wide behaviors of students would improve after Tier 1 implementation of the CW-FIT program. Because the use of exclusionary discipline practices is traditionally measured at the school-level, it was expected that little to no information would be provided surrounding the implementation of CW-FIT and reductions in the occurrence of teacher-delivered exclusionary discipline practices in the classroom.

Method

To address the present study's research questions, researchers conducted a comprehensive search of the literature surrounding the implementation and evaluation of the CW-FIT. A multi-gated approach was used to find articles, screen for inclusion eligibility, and code for information provided in the included studies. Independent reviewers replicated the search process to verify the reliability of initial findings. After screening and coding eligible articles, the final sample of included studies were evaluated to measure the quality of CW-FIT implementation and its effectiveness. To further quantify the effectiveness of Tier 1 components of the CW-FIT, a quantitative synthesis was conducted to quantify effects of single-case research design (SCRD) studies.

Search Procedures

The goal of the search process was to find all published studies and theses/dissertations that implemented the CW-FIT and evaluated its effects on class-wide student behavior. Therefore, researchers conducted a search of online databases (i.e., Academic Search Premier, Education Source, ERIC, ProQuest Dissertations and Theses, and PsycINFO) using the same keywords (i.e., CW-FIT OR Class wide function related Intervention* Team*) searched in both the Abstract (AB) and Title (TI) fields. We reported the yield of the initial search after removing duplicates, and how many articles were screened by title and abstract. Two independent reviewers conducted the same search process as the first author which resulted in 100% interrater reliability (IRR). The search was completed in September 2022.

Inclusion Criteria

Researchers screened the title, abstract, and when needed, full-text of non-duplicate articles found in the initial search. Then, we determined if the study met the following inclusion criteria:

1. Written in English
2. Conducted in the United States
3. Primary aim was to analyze implementation of Tier 1 components of the CW-FIT
4. Conducted in a K-12 school setting

5. Empirical study that quantitatively measured student behavior change
6. Classified as a thesis, dissertation, or peer-reviewed publication
7. Sufficient information provided to report or calculate effect size (ES) estimates
8. Behavioral data were reported at the classroom-level

Coding Procedures

To obtain information accurately and comprehensively from the included articles, we coded studies for demographic, procedural, and methodological characteristics. Additionally, in an effort to capture information related to study quality, a study quality coding protocol was developed based on quality indicators (QIs) from standards developed by the Council for Exceptional Children (CEC; Cook et al., 2015). Two independent reviewers conducted the same coding process as the first author which resulted in 90.68% IRR; where discrepancies occurred, discussions ensued until consensus was reached. A detailed coding scheme is available from the first author.

Study Characteristics

Information from these codes include study-level characteristics (e.g., authors, year, publication outlet). For example, it is necessary to code for publication year to determine if this literature base is outdated or modern. Furthermore, because the creation and evaluation of the CW-FIT was supported by a federal grant, it is important to know if the same groups of authors are publishing results on the CW-FIT. Studies were also coded to distinguish if researchers used an SCRD or group design (GD).

Demographic Characteristics

Descriptions within studies of participants, settings, and interventionists were also coded. This set of codes quantifies the level of homogeneity or heterogeneity that exists in the settings where the CW-FIT was implemented, evaluated, and disseminated. Similarly, these codes described the population of students, teachers, and interventionists who were exposed to the CW-FIT intervention program. The school type (e.g., public, private, charter) and educational setting (e.g., general education, special education, inclusive) were coded for each study to assess where the CW-FIT intervention was implemented. These codes provided information pertinent to knowing the common implementation and evaluation settings of the CW-FIT.

The race and ethnicity as well as gender of teachers and students were also coded. For teachers specifically, their education-level and years of experience were documented. For students specifically, their grade-level and disability type, if applicable, were both coded. For interventionists specifically, their role, education-level, and characteristics of their training in the CW-FIT were all coded. Information from these codes gave a clear understanding on whether certain subpopulations were more likely to experience the CW-FIT intervention program under empirical conditions.

Procedural Characteristics

Information from these codes included descriptions of the CW-FIT intervention procedures implemented. The focus of the present review, Tier 1, involves a group contingency where points are earned when all students on a team engage in the CW-FIT skills. Appendix A describes in greater detail the core components of the CW-FIT at Tier 1 (Wills et al., 2009).

Methodological Characteristics

Information from these codes included descriptions of the research methodology and design, implementation fidelity, and measures used in the included studies. From these codes, the overall methodological quality of the study was ascertained. This is critical in understanding the extent to which the knowledge generated to evaluate the effectiveness of the CW-FIT was accomplished using sound and appropriate research methods.

Quality of Included Studies

A unique aspect of the CEC’s standards is that it provides QIs for both SCRD and GD studies. Thus, to comprehensively assess the effectiveness of the CW-FIT, information on the overall quality of studies utilizing both SCRD and GD were coded. The CEC’s standards include recommendations on the amount and type of evidence that places a practice into one of their five classifications: (1) evidence-based practice, (2) potentially evidence-based practice, (3) practice with mixed effects, (4) insufficient evidence, and (5) practice with negative effects. Based on findings from the present systematic review, the CW-FIT program was placed into one of these classifications. The purpose of this is to demonstrate the extent to which the Tier 1 components of the CW-FIT were effective in managing the class-wide behaviors of students.

Analysis of Study Outcomes Included in Systematic Review

The present review included studies utilizing either SCRD or GD methodology. Different analyses were required to evaluate study outcomes from these two research designs (e.g., visual analysis, various ES metrics).

SCRD Studies

Visual Analysis. For SCRD studies, visual analysis is commonly used to evaluate the extent to which a functional relationship exists between variables. Although it is expected for all study authors to report visual analysis of their data, the first author conducted independent visual analyses of all included SCRD studies. Using the Visual Analysis Worksheet provided by Ledford et al. (2018), the first author systematically assessed the level, trend, variability, consistency, overlap, immediacy, and presence of a functional relationship based on the time series graphs provided.

Effect Size Estimation. In addition to visual analysis, it is possible to quantify the effects from SCRD studies via ES estimates. Unfortunately, there remains a lack of consensus in the literature on which ES estimates researchers should use, and all available ES estimates for this type of design have their limitations (Shadish et al., 2015). In order to estimate treatment effects, data were first extracted from the time-series graphs provided by authors via WebPlotDigitizer (Rohatgi 2012). Then, the SingleCaseES package by Pustejovsky and Swan (2019) from RStudio (version 3.5.2) was used to calculate ES estimates.

Tau/Tau-U. Tau/Tau-U are ES estimates that are commonly reported with studies using SCRD in large part because the Tau-U estimate takes into account serial dependency, or autocorrelation in baseline (Vannest & Ninci, 2015). To this end, an online calculator from Vannest et al. (2016) was used to assess baseline trend and determine if a correction (i.e., Tau-U) needs to be calculated instead of the standard Tau

ES estimate. Both Tau and Tau-U use a non-overlap technique which means these ES estimates quantify the amount of datapoints across phases that do not overlap. With this type of research design, the more datapoints that do not overlap the larger the effect. According to Parker et al. (2011), a Tau-U value ≤ .36 is in the 25th percentile, a value between .37 and .63 is in the 50th percentile, a value between .64 and .93 is in the 75th percentile, and a value between .94 and 1.00 is in the 90th percentile.

PAND. PAND was also calculated to quantify the effects of the CW-FIT. PAND focuses on the nonoverlap of data across phases by including all of the data to estimate treatment effectiveness (Lenz, 2013). Importantly, the PAND ES estimate addresses limitations of similar non-overlap ES estimates by not emphasizing a single data point in baseline. This limits the potential for ceiling or flooring effects that can prevent accurate evaluation of effects (Lenz, 2013). So, to interpret the magnitude of the quantitative effects from PAND, the percentile ranks from Parker et al., (2011) was used. According to Parker et al. (2011), a PAND value ≤ .20 is in the 10th percentile, a value between .20 and .38 is in the 25th percentile, a value between .38 and .64 is in the 50th percentile, a value between .64 and .86 is in the 75th percentile, and a value between .86 and 1.00 is in the 90th percentile.

GD Studies

Effect Size Estimation. For GD studies, Cohen's *d* ES estimates are frequently reported to estimate the magnitude of an effect an intervention had on an outcome (Durlak et al., 2009; Maher et al., 2013). Yet, there are limitations commonly cited in the literature about the likelihood of Cohen's *d* ES estimates to overestimate effectiveness in studies with smaller sample sizes and Hedge's *g* is recommended (Maher et al., 2013). An online ES calculator was used to calculate Hedge's *g* (Effect Size Calculator, 2020).

Hedge's g. Hedge's *g* is a commonly used ES estimate in group design studies that calculates standardized mean differences between groups. Using standardized estimates allows for the comparison of ES estimates across studies. The purpose of Hedge's *g* is to recognize the positive bias inherent in Cohen's *d* through pooling standard deviations (Durlak et al., 2009). Therefore, it is most appropriate to use Hedge's *g* when study sample sizes are smaller.

Analysis of Study Outcomes Included in Quantitative Synthesis

To understand the overall effects of Tier 1 components of the CW-FIT, researchers conducted a quantitative synthesis to combine findings from SCRD studies. To further understand *for whom* and *under what conditions* Tier 1 components of the CW-FIT are effective in improving the class-wide behavior of students, an effect size metric (i.e., Hedge's *g*) was calculated and aggregated across different study characteristics. In addition to an overall effect size, potential moderators were evaluated. Specifically, comparisons across treatment adherence (i.e., high or low) was conducted as exploratory.

All analyses for the quantitative synthesis were conducted in R using the scdhlm (Wang & Maxwell, 2015) and metafor packages (Schwarzer, 2022). The between-case standardized mean difference (BC-SMD) effect size (also known as "DHPS") provides a tool for comparing the results of SCRD studies to the results of GD studies (Maggin et al., 2017). There are technical limitations of the BC-SMD effect size that limits its application to only studies that have at least three individual participants and that use a treatment reversal, multiple baseline, or multiple probe design. These requirements are necessary because the BC-SMD effect size relies on quantifying variation in the outcome between-participants (Maggin et al., 2017). The estimates calculated for each AB phase contrast across classrooms were combined as an average to obtain a single effect size for each case (Pustejovsky & Ferron, 2017).

Eligibility Criteria for Quantitative Synthesis

Not all SCRD studies included in the systematic review met the necessary requirements to be included in the quantitative synthesis. Specifically, the following eligibility criteria needed to be met:

1. Included more than three classrooms in analyses.
2. Used a treatment reversal, multiple baseline, or multiple probe design.

A flow diagram for study inclusion is provided in Figure 1.2.

Of the 19 studies included in the systematic review, less than half (n = 7) met eligibility criteria and were included in the quantitative synthesis. The most common criteria that studies did not meet was that they only included a single classroom in their analyses. This highlights the difficulty of attempting to quantitatively synthesize results from SCRD studies. While the BC-SMD effect size measure requires three or more groups, it is not necessary for studies using SCRD to have more than one group because a single group can serve as its own control and does not require a control group condition like in GD studies.

Pooling Effect Sizes

Prior to pooling effect sizes, the heterogeneity of studies was assessed using Cochran's X^2-based Q-test and the I^2-test (Higgins & Thompson, 2002). The statistical model used to explain heterogeneity (or variation) within a quantitative synthesis can either be a fixed-effect or random-effects model. The fixed-effect model assumes that variation occurs only due to sampling error, not because there are actual differences in effect sizes across study samples. The random-effects model assumes that variation occurs because there are actual differences in effect sizes across study samples. For those with statistical heterogeneity (defined as $p \leq 0.10$ or $I^2 \geq 50\%$) the random effects model was used. The fixed effect model was used for studies without significant heterogeneity (defined as $p > 0.10$ or $I^2 < 50\%$) (Overton, 1998).

To pool effect sizes across all studies included in the quantitative synthesis, the BC-SMD ES estimate was calculated then the Hedge's *g* correction was applied which is consistent with recommendations regarding best practices for SCRD reporting (Shadish et al., 2014). Hedge's *g* was calculated using the following formula from Shadish et al. (2014)[1]:

$$g = \left(1 - \frac{3}{4v - 1}\right)\left(\frac{M_t - M_b}{S_p}\right)$$

A higher score is equated to a greater magnitude of difference or effect between groups. This ES estimate accounts for the positive bias associated with smaller sample sizes through pooling standard deviations as well as autocorrelation (Shadish et al., 2015). Thus, Hedge's *g* was used because it addresses

[1] Where *v* is an estimated degrees of freedom; M_t is the average of treatment observations; M_b is the average of baseline observations; S_p is the pooled standard deviation

common characteristics of studies using either a SCRD or a GD. For any test or model used when pooling effect sizes, $p < 0.05$ was considered statistically significant.

Publication Bias

Publication bias occurs when highly favorable or statistically significant findings are more likely to be published than findings with null or negative effects. Due to the popularity of results from both systematic reviews and meta-analyses being used to make high-stakes decisions related to educational legislation, policy, and practice, then it is imperative to determine the extent to which publication bias impacts those results (Ekholm & Chow, 2018).

To evaluate the presence of publication bias, a forest plot was created using information from all studies included in the quantitative synthesis. These studies were sorted on the forest plot according to their standard error. It is expected that if the overall effect estimate shifts to the right of the forest plot as an increased number of less precise studies are included then there exists publication bias (Ekholm & Chow, 2018).

To further evaluate the existence of publication bias and its influence on results, a funnel plot was created with information from the studies included in the quantitative synthesis. The funnel plot allowed for visual inspection of the existence of possible publication bias. It is expected for more precise studies (i.e., larger N with smaller standard error) to have effect sizes that are closer to the overall effect identified via the quantitative synthesis. However, if there are fewer studies located in the bottom left portion of the funnel plot than the bottom right portion of the funnel plot, then results from the quantitative synthesis are more influenced by publication bias (Ekholm & Chow, 2018).

Both forest and funnel plots are tests that rely on visual inspection to determine the existence of publication bias. Since there are no statistical criteria used to determine the severity of publication bias, the Egger's regression test was conducted to supplement findings from the funnel plot. The Egger's regression test "provides researchers with a statistical test for publication bias in which statistically significant results indicate the presence of bias" (Ekholm & Chow, 2018, p. 431). Specifically, it tests for asymmetry in the funnel plot.

Subgroup Analyses

To determine the extent to which heterogeneity in the overall pooled effect size estimate could be explained by scientific hypotheses, specific subgroup analyses were conducted. This included the extent to which different levels of treatment adherence across studies impacted results. In addition to obtaining the pooled effect for each subgroup, these estimates were also compared using a statistical test to determine significance (Borenstein & Higgins, 2013). As described above, the same criteria were used to determine if the fixed-effect or random-effects model was used. Plus, any test where $p < 0.05$ was considered statistically significant.

Results

Systematic Review

The electronic search initially yielded a total of 324 studies. After screening articles based on the inclusion criteria, then a final total of 19 studies were included in the systematic review. A flow diagram for study inclusion is provided in Figure 1.1. Researchers then coded these 19 studies to comprehensively document the characteristics of studies. The purpose of this systematic review was to understand for whom and under what conditions researchers evaluated the Tier 1 implementation of the CW-FIT.

Study Characteristics.

There was only one study classified as a thesis with the remaining 18 studies (94.7%) published in a peer-reviewed journal. A total of five studies (26.3%) were published in the *Journal of Positive Behavior Interventions* (JPBI). The primary purpose of 18 studies (94.7%) was to examine the effectiveness of the CW-FIT alone, with one study examining the effectiveness of the CW-FIT and the Good Behavior Game (i.e., Parikh, 2019). There were 10 studies (52.6%) published between 2011 and 2017, a total of 17 studies (89.5%) shared at least one co-author, and 16 studies (84.2%) utilized an SCRD. Table 1.1 provides more information on the characteristics that describe the included studies.

Demographic Characteristics.

Three studies did not report the grade-level of participating classrooms, but of the 16 studies (84.2%) that did, only one (6.3%) of them took place in a high school setting. Four studies (25.0%) took place in lower elementary, or grades K-2, classrooms, one study (6.3%) occurred in upper elementary, or grades 3-5, classrooms, and three studies (18.8%) took place in middle, or grades 6-8, classrooms. A total of seven studies (43.8%) included classrooms that took place in a mixture of lower and upper elementary as well as middle grade classrooms. A total of seven studies did not report what type of schools the study took place in, however, for the 12 studies (63.2%) that did report school type, nine studies (75.0%) occurred in public schools, followed by two studies (16.7%) in charter schools, and then one study (8.3%) in a private school. There were eight studies that did not report where the participating classrooms were located. However, for the 11 studies (57.9%) that did report, seven (63.6%) of the classrooms were located in an urban setting, with two (18.2%) in suburban settings, and one (9.1%) in multiple settings. Of the seven studies (36.8%) that reported free and reduced-price lunch (FRL), six aggregated their FRL estimate at the school-level and ranged from 47.6 percent to 94.0 percent. One study (i.e., Speight et al., 2021) reported an FRL estimate for the specific students included as participants in their study.

For educational settings, 10 studies (52.6%) occurred in regular education classrooms with a couple of studies (10.5%) occurring in special education classrooms as well as a single study (5.3%) occurring in an inclusive classroom. Two studies (10.5%) occurred in multiple educational settings. The educational subject describing when the CW-FIT was implemented varied across studies with seven (36.8%) reported multiple different educational subjects. However, specials such as art and music (26.3%), English Language Arts (15.8%), and math (10.5%) were all reported as educational subjects within which the CW-FIT was implemented. A couple of studies (10.5%) did not report during which educational subject the CW-FIT was implemented.

The average number of classrooms from the included studies was 3.4 with one study not reporting this information at the classroom-level but reported how many schools participated in the study (i.e., Wills et al., 2018). The number of student participants varied across studies with a range from 5 to 28 students per classroom. There were a couple of studies (13.3%) that did not report this information at the classroom-level but reported how many students there were at the school-level (i.e., Kamps et al., 2015b; Wills et al., 2014).

Students. While 10 studies (52.6%) did not report the gender of student participants, the studies that did report gender ranged from having 16.7 percent to 54.0 percent female participants. A total of seven studies (36.8%) did not report the race and ethnicity of participating students. Of the twelve studies (63.2%) that did report this information, the composition of White participating students ranged from 4 percent to 71.8 percent, the composition of Black participating students ranged from 0.0 percent to 70.0 percent, the composition of Latine participating students ranged from 0.0 percent to 87.0 percent, and the composition of Asian participating students ranged from 0.0 percent to 46.7 percent. Studies did not explicitly report having multiracial participating students, but the composition of participating students who identified as 'other' ranged from 0.0 percent to 7.1 percent. A total of 13 studies (68.4%) did not report if any participating students were English language learners (ELLs). However, of the 6 studies (31.6%) that did report this information, the composition of ELL participating students ranged from 2.7 percent to 62.0 percent. A total of 10 studies (52.6%) reported participating students having various identified or suspected disabilities. These disabilities included autism spectrum disorder (ASD), emotional behavioral disorder (EBD), speech or language impairment (SLI), learning disability (LD), other health impairment (OHI) due to attention-deficit/hyperactivity disorder (ADHD), and specific learning disability (SLD).

Teachers. Only three studies did not report the gender of participating teachers. Of the 16 studies (84.2%) that reported this information, the composition of participating teachers who were female ranged from 50 percent to 100 percent. A large proportion of studies (73.7%) did not report the race and ethnicity of participating teachers. Of the five studies that did report this information, most (80.0%) reported that 100.0% of their participating teachers were White. A single study (i.e., Speight et al., 2020) reported that two of their three participating teachers (67%) were White, and the remaining teacher (33%) was Black. A total of 16 studies (84.2%) reported teachers' years of experience. Of these studies, the amount of experience participating teachers had ranged from 1 year to 30 years. Eight studies (42.1%) did not report the educational level of participating teachers. Of those that did, three studies (27.3%) reported including teacher participants who had a master's degree and a couple of studies (10.5%) had teachers whose educational level was defined as 'other' (e.g., licensures, certifications). The remaining studies included multiple classroom teachers with varying educational levels that ranged from bachelor's to doctorate degrees.

Procedural Characteristics

In all studies, classroom teachers served as the interventionists who implemented the CW-FIT. For 11 studies (57.9%), general education teachers served as interventionists, with three studies (15.8%) reporting special education teachers as interventionists, and another four studies (21.1%) having teachers unspecified as interventionists. The remaining study (i.e., Speight et al., 2021) included both a general education teacher and a special education teacher as interventionists. Because all of the interventionists were teachers, the education level of those who implemented the CW-FIT match the participating teachers' education level previously described. For training on how to implement the CW-FIT, four studies (21.1%) did not explain the type of training teachers received. Of the 15 studies that did report this information, 12 (80.0%) involved the authors or developers of the CW-FIT delivering training to teachers, and the remaining three (20.0%) described using a CW-FIT coach to train teachers on delivering the intervention.

Appendix A describes Tier 1 implementation of the CW-FIT which comes from the comprehensive descriptions of CW-FIT implementation by Wills et al. (2009). Of the outlined procedures for Tier 1 implementation of the CW-FIT, five studies (26.3%) reported implementing all 12 components (i.e., Conklin et al., 2017; Hansen et al., 2017; Weeden et al., 2016; Wills et al., 2018; Wills et al., 2022). One study (5.3%) implemented almost all of the treatment components and only missed implementing one out of the 12 (i.e., Kamps et al., 2015a). A total of four studies (21.1%) reported implementing most, or 10 out of the 12, CW-FIT Tier 1 procedural components (i.e., Caldarella et al., 2015; Naylor et al., 2018; Nelson et al., 2018; Wills et al., 2014). A couple of studies (10.5%) reported implementing some, or nine out of the 12, CW-FIT procedures (i.e., Caldarella et al., 2017 and Kamps et al., 2011). Four studies (21.1%) reported implementing only a few, or eight out of 12, CW-FIT procedures (i.e., Hirsch et al., 2016;

Kamps et al., 2015b; Orr et al., 2020; Speight et al., 2021). The studies that reported the fewest CW-FIT procedures only reported implementing seven out of the 12 CW-FIT procedural components (i.e., Monson et al., 2020; Parikh, 2019; Speight et al., 2020). The most common CW-FIT Tier 1 procedures not reported implemented in studies were (1) whether teachers modeled examples and nonexamples of the skill, (2) whether students reinforced each other's appropriate behavior with attention and help, and (3) whether students received a reward for minimizing or ignoring attention to peers for inappropriate or problem behavior. All studies reported implementing the following CW-FIT procedures: reviewed the skill posters, recognized and rewarded students for appropriate behavior, implemented a group contingency, and gave teams points when they appropriately applied CW-FIT skills.

It is important to note that five studies (26.3%) incorporated CW-FIT Tier 1 procedures outside of the 12 outlined by Wills et al. (2009). Specifically, beyond the traditional three CW-FIT skills, students were taught either additional (i.e., Caldarella et al., 2017) or different skills (i.e., Monson et al., 2020), all instruction of CW-FIT sills and the social skills posters were translated from English to French (i.e., Hansen et al., 2017), instead of receiving a reward immediately after meeting the pre-determined goal, rewards were provided every second session for one classroom (i.e., Kamps et al., 2011), and finally, one teacher used an inaudible timer and did not have a point chart visible throughout the lesson which were both modifications from the original implementation of the CW-FIT (i.e., Nelson et al., 2018).

While only four studies (21.1%) reported monitoring generalization or maintenance of effects after CW-FIT implementation, all studies reported using rating scales to measure implementation fidelity data. A total of 17 studies (89.5%) reported how often implementation fidelity data were collected with a range from 17.0 percent to 100.0 percent of intervention sessions. Of the 18 studies (93.3%) that reported implementation fidelity data, treatment adherence ranged from 77.8 percent to 97.7 percent.

Methodological Characteristics

A total of 17 studies (89.5%) measured group on-task behavior as their outcome variable. One study measured both class-wide disruptive behaviors and compliance (i.e., Parikh, 2019) while the other study measured class-wide compliance (i.e., Conklin et al., 2017). All studies utilized systematic direct observations (SDOs) by external observers to measure their outcomes. Specifically, 15 studies (78.9%) reported using momentary time sampling and the remaining four studies (21.1%) used the Multi-Option Observation System for Experimental Studies (MOOSES).

For social validity, 14 studies (73.7%) measured the acceptability and feasibility of the CW-FIT for both interventionists (i.e., teachers) and recipients (i.e., students). Three studies (15.8%) only collected social validity data from the teachers who implemented the CW-FIT. Overall, the social validity data reported indicated that teachers and students from all 14 studies perceived the CW-FIT as a positive class-wide behavior management intervention.

All of the included studies focused on Tier 1 implementation of the CW-FIT to address class-wide misbehavior. Yet, none of the studies reported measuring change in the teacher-delivered exclusionary discipline practices in the classroom as an outcome measure. Instead, outcome measures were solely observations from external observers and did not include additional sources of information like documentation of class-wide exclusionary discipline practices.

Characteristics and Quality of SCRD Studies

As mentioned previously, most studies implemented an SCRD to evaluate the effects of the CW-FIT. More specifically, the most common type of SCRD utilized was withdrawal (36.8%) followed by multiple baseline design (31.6%). Cook et al. (2015) provided QIs for studies using an SCRD that included a range of characteristics to describe the requirements for high-quality implementation of this study design. For example, the number of demonstrations, sufficient baseline, adequate design to control for threats to internal validity, and inclusion of time series graphs are all aspects of the QIs Cook et al. (2015) provided that relate specifically to SCRD studies. The percentage of these QIs each SCRD study met are presented in Table 1.1. The most commonly missing QI in these studies were authors not reporting ES estimates (47.4%). One study using an SCRD did not include time-series graphs (i.e., Wills et al., 2022). All remaining SCRD studies included time-series graphs and reported having a baseline with three or more datapoints as well as demonstrated the effect at least three times.

Outcomes from SCRD Studies

Visual Analysis. There was a total of 35 time-series graphs available from the included studies because some studies included multiple graphs across participating classrooms. Thus, the Visual Analysis worksheet from Ledford et al. (2018) was completed for the 35 different time-series graphs. Data from a total of 34 graphs (97.1%) demonstrated consistent level within and between phases. Data from one graph did not establish a stable level prior to the introduction or withdrawal of the CW-FIT (i.e., Naylor et al., 2018). Data from all graphs demonstrated trends in the expected direction during baseline and intervention that made behavior change easy to detect as well as consistent change in trend across phases. Data from a total of 34 graphs (97.1%) included variability that did not interfere with the ability to observe level changes across phases. Data from one graph included unexpected variability that did interfere with the ability to observe level changes across phases (i.e., Naylor et al., 2018). Data from all graphs demonstrated high consistency within and between phases, minimal overlap across phases, and immediacy of effect with a discrepancy between the final and initial data points of each phase in the expected direction. A functional relation was present for all graphs because they all included at least three demonstrations of effect. Of these 35 graphs that demonstrated a functional relationship, 26 graphs (74.3%) yielded an extremely confident rating with nine graphs (25.7%) yielding a quite confident rating that the functional relationship was present. Additionally, it was determined that of these 35 graphs, 11 (31.4%) demonstrated medium effects and the remaining 24 (68.6%) demonstrated large effects.

Tau/Tau-U. Table 1.2 includes all of the Tau/Tau-U ES estimates and their relative effectiveness based on percentile ranks. Because one study needed to correct for baseline trend (i.e., Parikh, 2019), there is only one Tau-U ES estimate, and all the rest are calculated as Tau. All of these ES estimates were classified as at or above the 75th and 90th percentiles. Most (56.3%) Tau/Tau-U ES estimates from included SCRD were at or above the 90th percentile.

PAND. Table 1.2 includes all of the PAND ES estimates, and their relative effectiveness based on percentile ranks. Of the 16 studies using an SCRD, one study did not include adequate information to calculate PAND (i.e., Wills et al., 2022). The remaining SCRD studies all had PAND ES estimates classified at or above the 75th and 90th percentiles. Most (66.7%) PAND ES estimates were at or above the 90th percentile.

Characteristics and Quality of GD Studies

Two of the three GD studies (66.7%) were randomized controlled trials and used general linear mixed modeling (GLMM) as the data analysis technique. One study used a quasi-experimental design and conducted *ANOVAs* and *t* tests (i.e., Caldarella et al., 2015). Cook et al. (2015) provided QIs for studies using GD that included a range of characteristics that describe the requirements for high-quality execution of this study design. For example, explanations on how groups were assigned, amount of attrition, potential differential attrition, and appropriate ES estimates are all aspects of the QIs Cook et al. (2015) provided that relate specifically to GD studies. The percentage of these QIs each GD study met are presented in Table 1.1. The most commonly missing QI in these studies were attrition and differential attrition with no studies explaining their percentage of missing data or whether data were missing more or less frequently for certain groups. One study did not report a GD ES estimate (i.e., Kamps et al., 2015b). The other studies reported Cohen's *d* (i.e., Caldarella et al., 2015 and Wills et al., 2018). Despite the relatively smaller sample sizes, no GD studies reported Hedge's *g* ES estimates.

Outcomes from GD Studies

Hedge's g. Table 1.3 includes all of the calculated Hedge's *g* ES estimates for GD studies. All of the ES estimates indicated that a meaningful change occurred between treatment and control groups. The Hedge's *g* ES estimates obtained from Kamps et al. (2015b) and Wills et al. (2018) demonstrated a noticeable larger effect than what was obtained in the other included GD study. Still, the Hedge's *g* ES estimate obtained from Caldarella et al. (2015) indicates the CW-FIT had a medium effect in improving class-wide on-task behavior.

Quantitative Synthesis

Pooled Effect Sizes

Although a total of 19 studies were included in the systematic review, 12 of these studies were excluded from the quantitative synthesis (see Figure 1.2 for more details). The restricted maximum likelihood estimator (Viechtbauer, 2005) was used to calculate the heterogeneity variance. Additionally, Knapp-Hartung adjustments (Knapp & Hartung, 2003) were used to calculate the confidence interval around the pooled effect. The heterogeneity of variance analysis was statistically significant [$Q(6) = 30.11$, $p < 0.0001$, $I^2 = 78.6\%$; 95% CI: 74.87%, 98.16%] which indicated substantial heterogeneity of effects and supported use of the random-effects model to pool effect sizes.

Across the seven studies that were included in the quantitative synthesis, the pooled effect size, assuming a random-effects model, was large [$d = 3.94$, 95% CI: 2.44, 5.45; $g = 1.90$] which indicates that the implementation of the Tier 1 CW-FIT components had a large effect ($p < 0.0001$) in improving class-wide behavior. In addition to calculating the overall effect of the CW-FIT on class-wide behavior, the purpose of the quantitative synthesis was to determine the extent to which specific characteristics influences the magnitude of effects.

Publication Bias

Visual inspection of the forest plot (Figure 1.3) indicate that publication bias might influence the results of the quantitative synthesis. This is mainly because the effect size estimates seemed to somewhat increase based on the standard error, but this pattern was not always consistent. To further evaluate the potential presence of publication bias, a funnel plot was also created from the studies included in the quantitative synthesis (Figure 1.4). Visual inspection of the funnel plot further supports that publication bias did not significantly influence the results of the quantitative synthesis. This is mainly because the distribution of studies and their effects are not automatically related to the precision of study results. Due to these conflicting results, the Egger's test was also used to investigate publication bias by testing the asymmetry of the funnel plot (Figure 1.4). These results suggest there is substantial presence of asymmetry ($z = 3.28$, $p = .001$). Given the results from the forest plot and the Egger's test, the effects from the quantitative synthesis are likely to be impacted by publication bias.

Subgroup Analyses

The extent to which treatment adherence was considered ‘high’ or ‘low’ moderated the results. The three studies in the ‘high’ treatment adherence subgroup reported that the CW-FIT was implemented with integrity between 90 and 96% of sessions (i.e., Conklin et al., 2017; Speight et al., 2020; Wills et al., 2014). The three studies in the ‘low’ treatment adherence subgroup reported that the CW-FIT was implemented with integrity between 78 and 88% of sessions (i.e., Hansen et al., 2017; Kamps et al., 2011; Nelson et al., 2018). The presence of higher treatment adherence was found to have a significant negative effect on class-wide on-task behavior. The effect of treatment adherence on class-wide on-task behavior was estimated using a regression coefficient, with a value of -1.83 (95% CI [-3.46, -0.20], $z = -2.20$, $p < .05$).

Discussion

The CW-FIT is a classroom behavior management intervention that received extensive financial support from the federal government which allowed researchers to implement and study the CW-FIT in various contexts. Yet, this is the first systematic review to evaluate the effectiveness of Tier 1 implementation of the CW-FIT in improving class-wide behavior. The present study included 19 studies in the final review and comprehensively appraised the quality of studies as well as the effectiveness of the CW-FIT in improving student behavior. Overall, the included studies in the present review support the implementation of Tier 1 components of the CW-FIT in improving class-wide behavior. Results from the quantitative synthesis came from a total of 7 studies which support that the overall effect size for Tier 1 components of the CW-FIT to improve class-wide on-task behavior is very large.

When reviewing the results from the quantitative synthesis, it is important to consider for whom the Tier 1 implementation of the CW-FIT was very effective. All of the studies in the quantitative synthesis took place in either an elementary or middle school, were located in either a suburban or urban setting, and had students of color make-up 28 to 76 percent of their student enrollment. Unfortunately, this means that results from the present study offer minimal insight into the effectiveness of Tier 1 implementation of the CW-FIT in high schools, schools located in rural settings, or schools with specific racial and ethnic groups represented in their student body. This reflects a larger theme in the literature where secondary schools and high school students are rarely the focus of research (Castillo et al., 2022) compared to elementary and middle schools, schools located in rural settings are also rarely the focus of research (Thier & Beach, 2019) compared to schools in urban settings, and researchers tend to include White participants more so than any other race and ethnicity (Grapin & Fallon, 2022). Still, the positive effects found in this study for implementing Tier 1 components of the CW-FIT to improve class-wide on-task behavior support future research in this area with other school and student characteristics not reflected in the present quantitative synthesis.

There was inconsistent evidence that publication bias influenced the large effects observed in the present study. This was unexpected because only one non-peer-reviewed study was included in the quantitative synthesis (i.e., Parikh, 2019). Yet, use of the statistical test provided a more objective measure of detecting publication bias, compared to visual inspection, which supports that publication bias did influence the results. Findings from the subgroup analyses suggest that when studies reported lower treatment adherence (i.e., 78 to 88 percent) then they documented greater improvement in class-wide behavior compared to studies reported higher treatment adherence (i.e., 90 to 96 percent). While this finding is indeed unexpected, it underscores the importance of future research to identify the percentage of treatment adherence necessary to observe positive effects that is empirically evaluated.

A secondary aim of the present review was to evaluate the impact implementation of the CW-FIT had on the use of exclusionary discipline practices by teachers in their classrooms. Unfortunately, none of the included studies reported this information. Researchers did not measure or consider the extent to which improvements in class-wide student behavior could affect the occurrences of teacher-delivered exclusionary discipline practices such as ODRs.

CW-FIT as an Evidence-Based Practice

In the present review, standards from the CEC were used to appraise the quality of included studies. Based on the evidence provided in the present review, the CW-FIT is an evidence-based practice (Cook et al., 2015). This classification is based on the fact that a total of three methodologically sound group comparison studies were included in this review that demonstrated positive effects across a total of

321 participating teachers. Unfortunately, it is impossible to know how many students participated in the CW-FIT because this information was reported at the school-level and not the classroom-level.

Additionally, 16 methodologically sound SCRD studies were included in this review that demonstrated positive effects with approximately 190 teachers and 890 participating students. Furthermore, none of the included studies demonstrated neutral or negative effects after implementation of the CW-FIT. Overall, according to the CEC's standards (Cook et al., 2015), there is sufficient evidence in the present review to support the classification of Tier 1 components of the CW-FIT as an evidence-based practice.

Even though according to CEC quality standards, Tier 1 components of the CW-FIT can be classified as an evidence-based practice, it is important to note that the current evidence-base is comprised predominately of studies conducted by the original developers of the CW-FIT. Thus, when analyzing the methodology used and outcome variables evaluated, clear patterns emerge. While it is promising that the current research findings support the classification of Tier 1 components of the CW-FIT to be evidence-based, more replications by independent researchers implementing different research methods along with examining novel outcome measures are needed.

CW-FIT and Exclusionary Discipline

The attempt to evaluate the effects of Tier 1 components of the CW-FIT on class-wide exclusionary discipline practices resulted in an empty review. As mentioned earlier, most studies evaluating the use of exclusionary discipline practices do so at the school-level (Algozzine & Algozzine, 2007; Anyon et al., 2016; Bottiani et al., 2017; Fisher & Hennessy, 2016; Mitchell & Bradshaw, 2013). However, classroom environments are considered vulnerable decision points when it comes to whether or not exclusionary discipline practices are used (Smolkowski et al., 2016). Therefore, it is imperative that we know the extent to which practices like the CW-FIT influence the use of these discipline practices. Unfortunately, in the present review, none of the included studies reported outcomes from classroom exclusionary discipline practices.

Limitations and Future Directions

While findings from the present review and synthesis hold great potential for future implementation of the CW-FIT, there are multiple limitations that need to be considered. First, our criteria prevented the inclusion of studies conducted outside of the United States. This limits the generalizability of the positive findings associated with implementation of the CW-FIT to U.S. student populations. However, one of the included studies successfully translated CW-FIT intervention materials from English to French and positive effects were demonstrated in a dual language classroom (i.e., Hansen et al., 2017). We also limited the present review to only include studies that took place in K-12 settings. Once again, this limits the generalizability of our positive findings because we did not include studies that took place in pre-K or setting beyond K-12.

Although this review attempted to include unpublished literature to guard against the influence of publication bias, only one thesis was included in the final analyses. Additionally, the focus of the present review was on Tier 1 class-wide implementation of the CW-FIT, however, this classroom behavior management program is multi-level and includes interventions beyond Tier 1 that were not evaluated in the present review. Therefore, the findings associated in this study are specific to class-wide behaviors (i.e., not individual student behavior) as well as Tier 1 (i.e., interventions from Tiers 2 and 3 were not evaluated).

Because similar research teams implemented and evaluated the CW-FIT, findings from the present review are limited in their scope as they rely on a fairly homogeneous group of research studies. Additionally, it is plausible that data published in one study comes from the same dataset published in another study. Thus, it is unclear if the data analyzed in this review are independent of one another as completely unique datasets. Because we are unable to confirm independence of all data analyzed, this limitation must be considered when interpreting the positive findings of the present review.

Future systematic reviews evaluating the CW-FIT could build off the findings from the present review in multiple ways. For example, future reviews should focus on evaluating the effectiveness of all levels of the CW-FIT beyond Tier 1 implementation. Other researchers should also consider reviewing the effectiveness of the CW-FIT in additional school settings such as implementation in preschool classrooms.

To ensure the validity and generalizability of the positive findings reported in the present review, independent studies by researchers outside the original developers are needed. So, future researchers should implement and evaluate the CW-FIT in various settings and conditions. Similarly, future research in this area should focus on expanding the type of outcome variables used to measure the effectiveness of the CW-FIT. This is especially true for measuring the extent to which implementation of the CW-FIT results in less frequent use of classroom exclusionary discipline practices.

Implications for Practice

The purpose of the CW-FIT is to improve student behavior in the classroom, and results from the present review support the effectiveness of the CW-FIT in meeting this purpose. However, it is important to know for *whom* and under *what conditions* these positive effects occurred. Based on the studies reviewed, participating classrooms came from varying regions throughout the U.S. that included students with diverse identities including their race and ethnicity, disability status, and socioeconomic status. However, only one of the included studies implemented the CW-FIT in high-school setting and none implemented it in rural settings. So, it is unknown if the positive findings from this review would translate under those conditions.

Conclusion

Overall, findings from this review support the use of the multi-component behavior management intervention package, the CW-FIT, in improving class-wide on-task behavior. However, it is important to note that most studies included in this review consisted of overlapping authors and research teams, highlighting an apparent need for additional studies by independent researchers. The lack of reporting on the effects of CW-FIT on preventing or minimizing exclusionary discipline practices was not surprising but is still an important area of expansion for future work in this area. Overall, Tier 1 implementation of the CW-FIT did improve the class-wide behaviors of students who participated in the studies included in the present review and quantitative synthesis.

Chapter 3

Equitable Discipline: Positive Behavioral Interventions and Supports in the Classroom

Positive behavioral interventions and supports (PBIS) is an implementation framework commonly used by school professionals to address the adoption and application of school-wide practices to systematically teach and reinforce positive behavior. According to the Florida PBIS Project (2021), as of 2019, over 27,000 schools in the U.S. reported implementing PBIS. Decades of research have supported the use of PBIS in improving school climate (Bradshaw et al., 2009), increasing prosocial behaviors (Bradshaw et al., 2010), and reducing the occurrence of office discipline referrals (ODRs; Bradshaw et al., 2012; Flannery et al., 2014; Noltmeyer et al., 2019). However, research also suggests that the extent to which the core features of PBIS are implemented greatly varies across schools (Pinkelman et al., 2015) which is concerning because empirical evidence supports that higher implementation fidelity of PBIS results in better outcomes (Pas et al., 2019). Although implementation fidelity of PBIS is often determined by total scores on implementation fidelity measures, it is also possible to use subscale scores on these measures to identify the specific components, or critical elements, of PBIS that might be particularly influential for promoting positive school and student outcomes. Importantly, there continues to be racial discipline disparities in the use of exclusionary discipline practices in schools, despite overall high implementation fidelity of PBIS (Bradshaw et al., 2010; Hilberth & Slate, 2014; Vincent & Tobin, 2010). Consequently, there needs to be further investigation surrounding the extent to which certain components of PBIS might relate to improved equitable discipline outcomes.

To this end, there are multiple PBIS implementation fidelity measures available to investigate associations between certain components of PBIS and (in)equitable discipline outcomes. Analysis of these scales could be particularly beneficial in identifying potential levers for change. While some preliminary research has examined this issue (Barclay et al., 2022), there remains a dearth of peer-reviewed studies in the published literature that explicitly examine the extent to which certain PBIS components relate to (in)equitable discipline outcomes. Indeed, a review of published studies suggests that significant gaps exist examining the relationship between PBIS fidelity subscales and (in)equitable discipline outcomes. For example, Childs et al. (2016) examined how subscale scores from PBIS implementation fidelity measures associated with the occurrence of ODRs but did not examine racial discipline disparities. Relatedly, Gage et al. (2019) and both Heidelburg et al. (2022) examined how the *total* score from PBIS implementation fidelity measures associated with school- and student-level equitable discipline outcomes but did not examine subscales. The purpose of the present study is to expand on findings from Barclay et al. (2022) and Heidelburg et al. (2022) by using data from multiple states across the U.S. to determine the potential influence of certain components of PBIS, as measured by the Benchmarks of Quality (BoQ), in reducing the occurrence of racial discipline disparities in the use of exclusionary discipline, particularly ODRs.

Measuring PBIS Implementation Fidelity

To ensure critical practices from PBIS are implemented as intended, there exist various PBIS implementation fidelity measures for leadership teams in schools to use, including: the Schoolwide Evaluation Tool (SET), Self-Assessment Survey (SAS), Tiered Fidelity Inventory (TFI), and the BoQ. These measures typically rely on a combination of self-report, direct observation, reviewing permanent products, and consulting with a PBIS coach or facilitator supporting the larger school PBIS leadership team. The SET is one of the earliest implementation fidelity measures empirically supported for school leadership teams to use to make decisions surrounding improving PBIS implementation (Horner et al., 2004; Pas et al., 2019). The subscales on the SET include expectations defined, expectations taught, reward system, violation system, monitoring and evaluation, management, and district support. The SAS is a fully self-report measure that includes four hypothesized factors (i.e., school-wide systems, nonclassroom setting systems, classroom systems, and individual student systems; Solomon et al., 2015). The TFI is unique because it measures implementation fidelity across the three tiers associated with PBIS implementation (McIntosh et al., 2017).

While there are many measures to choose from, it is important for leadership teams to measure implementation fidelity of PBIS because lower implementation fidelity is associated with poorer outcomes at both the student-level (Pas et al., 2019) and the school-level (James et al., 2019). However, the measure leadership teams use will influence the type of data they collect regarding PBIS implementation. For instance, the SET offers information surrounding broader implementation of PBIS related to systems, whereas results from the SAS indicate which specific systems stakeholders would like to prioritize for improved implementation. The TFI serves as a summative measure indicating whether or not certain components at each tier of PBIS are implemented. Despite the utility of these measures and the evidence supporting their use, none of them except the BoQ provide nuanced information relating to systems (e.g., classrooms) and practices (e.g., data entry) that are pivotal to successful implementation of PBIS in schools. Thus, the focus of the present study is on the BoQ because it addresses the limitations of other PBIS implementation fidelity measures.

Benchmarks of Quality

The BoQ is one of the PBIS implementation fidelity measures that school PBIS leadership teams complete to evaluate implementation of Tier 1 PBIS practices. The BoQ is a self-evaluation tool intended to guide both initial implementation and sustained use of PBIS at Tier 1. As originally developed, the BoQ was a 53-item measure that consisted of 10 subscales or critical elements which included *PBIS Team*, *Faculty Commitment*, *Effective Discipline Procedures*, *Data Entry*, *Expectations and Rules*, *Reward System*, *Lesson Plans*, *Implementation Plans*, *Crisis Plans*, and *Evaluation* (Kincaid et al., 2005). To develop the measure, items that were identified as critical elements of PBIS and subsequently incorporated within the FLPBIS training manual were used (Cohen et al., 2007). A number of psychometric properties for the original BoQ were evaluated by Cohen et al. (2007) including Cronbach's coefficient alpha (0.43 to 0.87), test-retest reliability (0.63 to 0.93), interrater reliability (average of 89%), and concurrent validity (with correlations between scores on the BoQ and the SET at 0.51, which was statistically significant). These psychometric properties described by Cohen et al. (2007) are often cited in the literature as evidence supporting the use of this measure when evaluating school and student outcomes.

However, it is important to note that the original BoQ (Cohen et al., 2007) was revised in 2011 to replace the *Crisis Plans* critical element with the *Classroom Systems* critical element (hereafter described as the BoQ Revised; BoQ-R). Due to the influence of classroom PBIS implementation in reaching an overall high level of PBIS implementation fidelity in schools, items from the *Crisis Plan* critical element were replaced with seven new items to create the *Classroom Systems* critical element (Childs et al., 2011). This is a unique addition to the BoQ-R over other PBIS implementation fidelity measures because the *Classroom Systems* critical element consists of uniquely developed items dedicated to monitoring classroom systems within a Tier 1 PBIS system. Compared to the BoQ-R, other PBIS implementation fidelity measures that directly assess classroom PBIS systems are either too narrow (i.e., TFI) or too broad (i.e., SAS). Specifically, the TFI only includes a single item to measure classroom-level PBIS implementation compared to the seven-items that comprise the *Classroom Systems* critical element on the BoQ-R. With regard to the SAS, all staff members in a school are expected to complete the measure. Consequently, there is not a need for everyone to reach consensus because results from the SAS are not intended to evaluate effectiveness of a PBIS system. Instead, results from the SAS are used to identify areas of improvement within a PBIS system. Conversely, the overall BoQ-R measure was developed to assess the extent to which different PBIS components are implemented. Results from the BoQ-R are used to evaluate sustained PBIS implementation, and it is recommended that an external PBIS coach or facilitator assist teams with accurately reporting the implementation fidelity of their Tier 1 system.

Psychometric Evidence Supporting the Use of the BoQ-R

Despite the potential of the BoQ-R to evaluate and monitor classroom implementation of PBIS practices, there are limited studies empirically evaluating the psychometric properties of the BoQ-R measure (Barclay et al., 2022; Childs et al., 2011). Among studies that have used the BoQ-R, it was used primarily to explore the association between overall PBIS implementation, as measured by the total score, and student-level and school-level outcomes. For example, researchers found a direct association between higher implementation fidelity, as measured by the BoQ-R, and fewer suspensions (Gage et al., 2018; Gage et al., 2019). In addition to an overall reduction in suspensions, Gage et al. (2019) found that both Black students and students with disabilities experienced significantly fewer suspensions in schools with higher implementation fidelity. Furthermore, different researchers reported that the BoQ-R scores from the *Classroom Systems* critical element predicted student discipline outcomes from school-level reports (Childs et al., 2016). Specifically, Childs et al. (2016) used multiple measures of school-wide discipline data (i.e., number of ODRs, days of in-school suspensions, and days of out-of-school suspensions for each school) to capture the extent to which changes in BoQ-R scores across years influence these school-level discipline outcomes. These researchers found that only two subscales from the BoQ-R significantly predicted a decrease in frequency of discipline outcomes (i.e., *Data Entry* and *Classroom Systems*).

Despite these promising results, it is important to note that inconsistent findings have also been reported exploring the relation between the BoQ-R and target outcomes. For example, Pas and Bradshaw (2012) found that total scores from the BoQ-R did not significantly impact student outcomes (i.e., math, reading, and truancy). Similarly, other researchers found that total BoQ-R scores did not predict different growth trajectories amongst ODRs and suspensions across high and low implementing schools (Childs et al., 2016; Kim et al., 2018). In light of these inconsistent findings, additional research is needed regarding the BoQ-R generally, but also specifically examining how critical elements relate to (in)equitable discipline outcomes.

To date, only two studies examined the relation between BoQ-R scores and (in)equitable discipline outcomes as measured by ODRs. Heidelburg et al. (2022) focused on evaluating PBIS implementation fidelity, as measured by the BoQ-R, in predicting the occurrence of ODRs for Black students. These researchers included six elementary (i.e., Kindergarten through 6^{th}) schools, three middle (i.e., 5^{th} – 8^{th}) schools, and 17 mixed grade-level (i.e., Kindergarten through 8^{th}) schools from a single school district. Findings resulted in a non-statistically significant relationship between total BoQ-R scores and the total number of ODRs received by Black students. These findings were not replicated in another study examining the relation between BoQ-R scores and discipline outcomes. In another study, Barclay et al. (2022) included a sample of 322 elementary, middle, and high schools throughout the state of Florida and found that higher scores from the *Expectations and Rules, Reward System*, and *Classroom Systems* critical elements on the BoQ-R were significantly related to lower overall risk of students receiving ODRs and experiencing suspensions, but this did not always mitigate racial discipline disparities (indicated by the risk index). For instance, scores for the *Expectations and Rules* and *Classroom Systems* critical elements did not have a significant association with the risk students from certain racial and ethnic groups (i.e., Black and Hispanic/Latine) faced in receiving an ODR or suspension. Yet, scores from the *Reward Systems* critical element did associate with more equitable discipline outcomes for Black students, in that scores were associated with reduced risk of suspension. In light of these mixed findings, further research is needed to examine these patterns within a more representative sample. Further, Barclay et al. (2022) included five critical elements (i.e., *Data Entry*, *Expectations and Rules*, *Reward System*, *Lesson Plans*, and *Classroom Systems*) and only focused on two specific racial and ethnic groups (i.e., Black and Hispanic/Latine). Thus, their analysis provided a relatively limited understanding of racial discipline disparities in the use of exclusionary discipline practices.

Due to the widespread implementation of PBIS across the U.S., and the persistent racial discipline disparities amongst Indigenous/Native students, it is necessary for future studies with samples representing multiple states with schools implementing PBIS to evaluate the association between scores from critical elements on the BoQ-R and (in)equitable discipline outcomes across various student racial and ethnic groups. The current study aims to address limitations in both Heidelburg et al. (2022) and Barclay et al. (2022) studies by including data from schools across the United States as well as comprehensively evaluating the relation between BoQ-R scores and discipline outcomes for students identifying as Asian, Black, Hispanic/Latine, multiracial, Indigenous/Native American, Pacific Islander, or White.

Current Study

To address the lack of empirical studies evaluating the association between core components of PBIS implementation and equitable discipline outcomes, the current study expanded on the findings from Heidelburg et al. (2022) and Barclay et al. (2022) by using data from multiple states across the U.S., including results from more diverse racial and ethnic groups, and evaluating all of the subscales on the BoQ-R. The following research questions guided the analyses:

1. What evidence is there for structural validity of the BoQ-R in a national sample?
2. To what extent does implementation fidelity of PBIS (as measured by the total score) from the BoQ-R as well as implementation fidelity of PBIS components (as measured by the specific

critical elements) from the BoQ-R (e.g., *Classroom Systems*), relate to a decrease in the average number of ODRs given to specific racial and ethnic groups within a sample of U.S. schools?

3. To what extent does implementation fidelity of PBIS (as measured by the total score) from the BoQ-R as well as implementation fidelity of PBIS components (as measured by the specific critical elements) from the BoQ-R (e.g., *Classroom Systems*), relate to greater equitable school discipline outcomes (i.e., lower risk ratios) across specific racial and ethnic groups within a sample of U.S. schools?

Method

Data Source

This study leveraged a large, national dataset maintained by the Educational and Community Supports (ECS) research unit within the College of Education at the University of Oregon. As a research unit housing the OSEP Technical Assistance Center on PBIS, ECS maintains national datasets that are available for limited use by external users under specific purposes. Specifically, ECS maintains data from schools in the U.S. that a) used School-Wide Information System (SWIS) in the most recent academic year of data with corresponding National Center for Education Statistics (NCES) information available (i.e., 2017-2018) and b) agreed to have their data used for research purposes. All data were reported at the school-level.

Measures

The dataset included various measures that researchers used in the regression models as covariates, predictors, or outcome variables. Covariates included in each model were selected to address variables that could potentially confound results. For example, it is plausible that the number of ODRs reported by schools increases as the number of students increase. Also, the percentage of students receiving FRL illustrates the unique needs that come with students from different socioeconomic backgrounds. Student needs vary over time which is reflected in the different experiences for students enrolled in an elementary, middle, or high school. The number of male students enrolled in each school was also considered as an important covariate to include due to a historical pattern in the literature that male students are more likely to experience exclusionary discipline from school (Office for Civil Rights, U.S. Department of Education, 2002; Petras et al., 2011). These covariates align with patterns from previous research in the literature evaluating the BoQ-R and discipline outcomes. Specifically, researchers planned to account for student enrollment by race and ethnicity (Barclay et al., 2022; Gage et al., 2018; Gage et al., 2019; Heidelberg et al., 2022; Kim et al., 2018), total student enrollment (Barclay et al., 2022; Childs et al., 2016; Gage et al., 2018), percentage of students receiving FRL (Gage et al., 2018; Gage et al., 2019; Kim et al., 2018), and grade-level (Barclay et al., 2022; Childs et al., 2016; Gage et al., 2018; Gage et al., 2019; Heidelberg et al., 2022; Kim et al., 2018). For predictors, total BoQ-R scores as well as scores from each of the BoQ-R subscales were included across models. Lastly, for outcome variables, both major and minor ODRs were included, such that researchers calculated average ODRs per racial and ethnic group and risk ratios for certain racial and ethnic groups and were included in relevant models. Researchers calculated average ODRs per racial and ethnic group for each school by dividing the number of ODRs per racial and ethnic group by the number of students in that racial and ethnic group (e.g., Average Asian ODRs = # of Asian ODRs / # of Asian students enrolled). Researchers calculated risk ratios for each school by using White students as the reference group. This means that risk ratios were calculated by dividing the average number of ODRs per student from the target racial and ethnic group by the average number of ODRs per White student (e.g., Risk Ratio for Asian ODRs = Average # of ODRs per Asian student / Average # of ODRs per White student).

Preliminary Analyses

Descriptive analyses were conducted to understand various demographic characteristics of the included sample. For example, demographic characteristics (e.g., grade level, urbanicity), school characteristics (e.g., free or reduced-price lunch; FRL), student characteristics (e.g., gender, race and ethnicity), ODR characteristics (e.g., average ODRs received by racial and ethnic groups), and PBIS implementation characteristics (e.g., total scores from the BoQ-R) were all reported. Next, preliminary data screening included the evaluation of statistical assumptions, like the identification of outliers, distribution of variables, and missing data. Researchers determined whether data were missing completely at random

(MCAR), missing at random (MAR), or missing not at random (MNAR). To test for MCAR, researchers conducted Little's test via SPSS (Li, 2013).

Measurement Model: Confirmatory Factor Analysis (CFA)

The CFA addressed our first research question by evaluating the extent to which the factor structure of the BoQ-R aligns with the structure hypothesized by its developers (Childs et al., 2011). The primary aim of this analysis was to confirm the internal structure of the BoQ-R, including evidence of both reliability and validity of scores, in the current sample prior to using specific critical elements as variables in subsequent regression analyses. This aligns with the argument-based approach to validation (Cook et al., 2015; Kane, 2016) with results from the CFA informing the validity argument in terms of the appropriateness of the proposed interpretations and uses of scores from the BoQ-R. This is especially important with the proposed use of critical element scores from the BoQ-R when most previous research studies only used total scores from PBIS implementation fidelity measures.

To assess goodness of fit within the CFA, multiple fit indices were used. Specifically, the Chi-Square test, comparative fit index (CFI), Tucker Lewis index (TLI), Standardized Root Mean Squared Error (SRMR), and the Root Mean Square Error of Approximation (RMSEA) were all calculated. For the Chi-Square test, a good model fit would yield an insignificant result at the 0.05 p-value threshold. For the both the CFI and TLI, a value greater than or equal to 0.95 is considered an indicator of good fit. For the SRMR, a value of less than 0.08 and for the RMSEA a value of less than 0.07 are both considered indicators of good fit. While all fit indices and their cut-off values have their limitations, these are commonly used in the literature (Hooper et al., 2008).

The estimation method used for the CFA was weighted least square mean and variance adjusted (WLSMV) because this is appropriate for ordinal data (Beauducel & Herzberg, 2006). The WLSMV estimation method requires no assumptions regarding the distribution of observed ordinal data, but it does assume a normal distribution of latent continuous data (Li, 2016). For the final model, internal consistency estimates were calculated using Cronbach's alpha to determine the extent to which items from each subscale in the BoQ-R reliably measured the same construct. Cronbach's alpha estimates range from 0 to 1 with higher values indicating greater internal consistency across items.

Multiple Linear Regression

To address the remaining research questions, two parallel models per outcome variable were built apriori on across racial and ethnic groups with relevant covariates. The specific outcome variables included were (1) the average number of ODRs per racial and ethnic group and (2) the risk ratios for receiving ODRs per racial and ethnic group. It is important to note that the models including risk ratios used the White racial and ethnic group as a comparison for all remaining racial and ethnic groups. The parallel models also included either (1) the modified total BoQ-R score or (2) subscale scores from the BoQ-R as predictors. This resulted in a total of 4 models per racial and ethnic group (excluding the White racial and ethnic group which had a total of 2 models). The purpose of all of these models was to capture the relation between implementation of PBIS and exclusionary discipline. Thus, if the qualifying assumptions were met, then researchers used the multiple linear regression in R via the function, lm, to estimate unstandardized regression coefficients, adjusted R-squared values, and p-values.

Assumptions.

Prior to conducting multiple linear regression, it was crucial to evaluate the extent to which the data from each model in the present study met the following assumptions: (1) linearity, (2) independence, (3) homoscedasticity, (4) normality. These assumptions were tested using various statistical packages in R. For linearity, a correlation coefficient was obtained between each predictor and outcome variable. Additionally, residuals were plotted against predicted values to assess the linearity and constant variance assumptions. To evaluate the presence of independence and homoscedasticity more objectively, both the Durbin-Watson and Breusch-Pagan statistical tests were conducted, respectively. Normality of residuals was assessed using Q-Q plots and the Shapiro-Wilk test, and the skewness and kurtosis of the residuals were also evaluated. Multicollinearity was evaluated using variance inflation factor (VIF) values.

Missing Data

If there was evidence that missing data in the present study were not MCAR, then researchers used the expectation-maximization (EM) estimation through multiple imputation (MI) an iterative process to provide estimates across the variables with missing data. The use of MI is a recommended approach for handling missing data when data are not MCAR (Peugh & Enders, 2004). The presence of more than one variable with missing data supports the use of fully conditional specification (FCS), or multivariate imputation by chained equations (MICE), via the mice package in R (Buuren & Groothuis-Oudshoorn, 2011). While a small number of iterations (i.e., between 10 and 20 imputations) is considered sufficient when using the mice package (Buuren & Groothuis-Oudshoorn, 2011), researchers also used the 'howManyImputations' package in R to determine the recommended number of imputations needed to obtain more precise standard error estimates based on a summary of the imputation variation (i.e., the coefficient of variation; von Hippel, 2020). To ensure stability of results obtained through the MI process, a value of .05 was used for the coefficient of variation so that when data were re-imputed, the estimated standard errors would not vary by more than this value. Thus, researchers used the two-stage procedure described by von Hippel (2020) to conduct sensitivity analysis for the MI data: (1) obtain estimates using a smaller number of datasets (i.e., 20) and (2) obtain estimates using the number of datasets recommended by the 'howManyImputations' package. To determine the number of imputed datasets to include in the final model, researchers compared the standard error estimates and the Fraction of Missing Information (FMI) values from the two-stage procedure. Then, results from all of the imputed datasets were combined and pooled to obtained parameter estimates and standard errors. The pooled estimates accounted for the uncertainty introduced by missing data to improve accuracy of findings.

Results

School Characteristics.

All of the schools (N = 333) included in analyses belonged to public school districts. Four of these schools were also designated as charter, nine were designated as alternative, and only five schools were designated as magnets. As designated by the U.S. census bureau, the geographical locations of these schools represented the four regions of the U.S. (i.e., Northeast, Midwest, South, and West). Specifically, schools came from a total of 18 states with three states in the Northeast region, six states in the Midwest region, three states in the South region, and six states in the West region of the U.S. (Geographic Areas Reference Manual, 2022). To determine the percentage of schools located in a suburban, urban, or rural geographical location, researchers also analyzed additional information related to where schools were geographically located. Of the 315 schools that reported this information, 59.68% (n = 188) were schools located in a suburban geographical location, 15.56% (n = 49) were schools located in an urban geographical location, and 24.76% (n = 78) were schools located in a rural geographical location.

The primary grade level each school served included elementary (n = 243), middle (n = 59), and high schools (n = 31). A select number of schools (n = 59) also reported serving pre-K students. Of the 313 schools that reported whether or not they received federal funding from the Title 1 program, 76.68% (n = 240) reported receiving Title 1 funds. While some schools did not report the number of students receiving free or reduced-price lunch (FRL), of those that did (n = 284), students receiving FRL ranged from 1.83% to 100% with an average of 48.6% of students receiving FRL across schools.

For student enrollment, of the 318 schools that reported this information, student enrollment ranged from 19 to 2,491 students with an average of 534 students enrolled in each school. The majority of schools (86.79%; n = 289) reported the number of students of color and White students in their population, which included disaggregating this information into various racial and ethnic groups (i.e., Asian, Black, Hispanic/Latine, Multiracial, Indigenous/Native American, Pacific Islander, White). Refer to Table 2.1 for more information about the student enrollment and school demographics.

Office Discipline Referral Characteristics.

All schools (N = 333) included in analyses reported a count of the total ODRs documented that ranged from 35 to 4,720 ODRs. An average of 556 ODRs per school were documented in the 2017-2018 school year. Of the schools that reported student enrollment, the number of ODRs per student ranged from 0.03 to 12.52 with an average of 1.84 ODRs per student.

Average Number of ODRs per Racial and Ethnic Group.

The total and average number of ODRs received by students from certain racial and ethnic groups depended on the number of schools that disaggregated information related to the race and ethnicity of students enrolled in their school. Therefore, it should be noted that these analyses are restricted to schools that reported ODR data disaggregated by the referent group. Refer to Table 2.1 for more information about the minimum, maximum, and average number of ODRs per racial and ethnic group.

Risk Ratios for ODRs Received per Racial and Ethnic Group.

The risk ratios for receiving ODRs by students from certain racial and ethnic groups compared to students from the White racial and ethnic group depended on the number of schools that disaggregated information related to the race and ethnicity of students enrolled in their school. Therefore, it should be noted that these analyses are restricted to schools that reported risk ratios disaggregated by the referent group. A total of 235 schools reported that the risk ratio for Asian students receiving an ODR compared to White students ranged from 0 to 7.29 with an average of Asian students being at a lower risk (i.e., 0.40) of receiving an ODR compared to White students in each school. A total of 264 schools reported that the risk ratio for Black students receiving an ODR compared to White students ranged from 0 to 133.64 with an average of Black students being at a higher risk (i.e., 2.79) of receiving an ODR compared to White students in each school. A total of 266 schools reported that the risk ratio for Hispanic/Latine students receiving an ODR compared to White students ranged from 0 to 16.67 with an average of Hispanic/Latine students being at a higher risk (i.e., 1.44) of receiving an ODR compared to White students in each school. A total of 184 schools reported that the risk ratio for Indigenous/Native American students receiving an ODR compared to White students ranged from 0 to 78.09 with an average of Indigenous/Native American students being at a higher risk (i.e., 1.09) of receiving an ODR compared to White students in each school. A total of 234 schools reported that the risk ratio for Multiracial students receiving an ODR compared to White students ranged from 0 to 102.82 with an average of Multiracial students being at a higher risk (i.e., 1.02) of receiving an ODR compared to White students in each school. A total of 113 schools reported that the risk ratio for Pacific Islander students receiving an ODR compared to White students ranged from 0 to 21.05 with an average of Pacific Islander students being at an equal risk (i.e., 1.00) of receiving an ODR compared to White students in each school. All of these results are further illustrated via histograms (Figure 2.1 through Figure 2.6).

PBIS Implementation Fidelity as Measured by the Benchmarks of Quality-Revised.

All participating schools (N = 333) reported total scores on the BoQ-R as well as total scores for each critical element. For the total BoQ-R score, schools ranged from scoring a 32 to 107 on the BoQ-R (M = 89.66, SD = 14.74). The most commonly reported total score on the BoQ-R was 102 (i.e., 6.0% of schools). Schools reported implementing all items from the following critical elements: *PBIS Team* (60.7% of schools), *Faculty Commitment* (29.1% of schools), *Effective Discipline Procedures* (54.7% of schools), *Data Entry* (36.0% of schools), *Expectations and Rules* (62.8% of schools), *Lesson Plans* (27.3% of schools), *Classroom Systems* (39.6% of schools), and *Evaluation* (28.2% of schools). Out of the 16 points that make up the total implementation fidelity score for the *Reward System* critical element, 19.5% (n = 65) of schools reported a score of either 14 or 15. Out of the total 13 points that make up the total implementation fidelity score for the *Implementation Plans* critical element, 18.0% (n = 60) of schools reported a score of either 12 or 13. There were three critical elements that a few schools reported a total implementation fidelity score of zero suggesting that these critical elements were not implemented at all (i.e., *Faculty Commitment*, *Lesson Plans*, and *Implementation Plans*). The critical element that the lowest number of schools reported implementing fully was *Reward System* (15.0%) and the critical element that the highest number of schools reported implementing fully was *Expectations and Rules* (62.8%).

Confirmatory Factor Analysis (CFA)

First, a baseline model was fit to the data, consisting of each item on the BoQ-R and the accompanying subscales (Model 1). The fit indices for Model 1 did not provide evidence to support strong model fit ($\chi 2$ = 2213.77, df = 1315, p < .001; CFI = 0.83; TLI = 0.82; SRMR = 0.17; RMSEA = 0.06). Therefore, modifications to the measurement model (Figure 2.1) were made based on evaluation of modification indices. Specifically, modification indices were evaluated using the expected parameter change (EPC) which measured the change in Model 1's chi-square test statistic. If the modification indices were greater than or equal to 10 then the parameter was removed (Wang & Wang, 2019). The only parameter that met this criterion was item eight from the BoQ-R which is part of the *Effective Discipline Procedures* critical element. Since the item asked whether or not respondents had a “discipline process [that] includes documentation procedures” by default, by using SWIS (and thus being included in the dataset), the response to this item was affirmative for all schools included. This is because all schools in the present study had access to SWIS which means that they all had documentation procedures in place. After removing item eight, the overall fit of the revised measurement model did support strong evidence for structural validity of the BoQ-R with a modified total score in a national sample of schools using SWIS ($\chi 2 = 1491.20$, df = 157, p = .000; CFI = 0.95; TLI = 0.95; SRMR = 0.11; RMSEA = 0.03). Thus, total scores from the BoQ-R (without item eight) were used in regression models to test the relation between overall PBIS implementation and exclusionary discipline.

In addition to evaluating the structural validity of the BoQ-R, researchers also evaluated the internal consistency of each subscale using Cronbach’s alpha. While there are no universally accepted cut-offs, a Cronbach’s alpha coefficient greater than or equal to 0.70 is often deemed acceptable (Taber, 2018). Researchers considered this criterion and determined that BoQ-R subscales with a 95% confidence interval that included a Cronbach’s alpha value greater than or equal to 0.71 would be included in subsequent analyses. Across the 10 subscales, all except two met this criterion (i.e., *PBIS Team* and *Faculty Commitment*). The Cronbach’s alpha coefficients and 95% confidence intervals for each BoQ-R subscale were as follows: *PBIS Team* ($\alpha = .31$, 95% CI: [.13, .46]), *Faculty Commitment* ($\alpha = .64$, 95% CI: [.56, .70]), *Effective Discipline Procedures* (without item 8; $\alpha = .75$, 95% CI: [.68, .80]), *Data Entry* ($\alpha = .69$, 95% CI: [.62, .74]), *Expectations and Rules* ($\alpha = .73$, 95% CI: [.64 , .79]), *Reward System* ($\alpha = .76$, 95% CI: [.70, .80]), *Lesson Plans* ($\alpha = .77$, 95% CI: [.71, .81]), *Implementation Plans* ($\alpha = .79$, 95% CI: [.74, .83]), *Classroom Systems* ($\alpha = .83$, 95% CI: [.78, .86]), and *Evaluation* ($\alpha = .74$, 95% CI: [.69, .79]). Although most of the BoQ-R subscales yielded results that supported the reliability of their items, the total scores from the two subscales that demonstrated low reliability in their items were not used in regression models to test the relation between specific PBIS implementation components and exclusionary discipline.

Multiple Linear Regression

For each research question, all assumptions were tested prior to conducting multiple linear regression analyses. Then, statistically significant findings were reported across each racial and ethnic group.

Missing Data

There were no missing data from the information used as independent variables (i.e., BoQ-R scores). This is because the national dataset obtained from the ECS required users to complete the full BoQ-R measure prior to submitting their data for use for research purposes. However, there were missing data from values used in outcome variables.

Not all schools reported their ODR data disaggregated by racial and ethnic group. Information about the average ODRs received by Pacific Islander students had the highest percentage of missing data (66.1%), followed by the average ODRs received by Indigenous/Native American students (44.4%). The average ODRs received by Multiracial students (29.1%) and Asian students (28.5%) both had a similar percentage of missing data. The racial and ethnic groups with the lowest percentages of missing data were Black students (19.5%), Hispanic/Latine students (18.6%), and White students (15.0%). Because all variables had missing data for more than 5% of cases, this was evidence to support that the missing data do not meet the assumption of MCAR. Results from the Little's MCAR test in SPSS, (X^2(494, N = 333) = 814.83, p = .000) provided additional evidence supporting that these missing data were not MCAR. Due to the evidence supporting that the missing data in the present study are not MCAR, researchers conducted the MI process described earlier. The number of imputations included in these results ranged from 20 to 181 imputed datasets, with all values pooled across these datasets to obtain the estimates used in the final models.

Research Question 2: Overall PBIS Implementation and Occurrence of Exclusionary Discipline

Assumptions for models using (1) the modified BoQ-R total score as a predictor and (2) average ODRs per racial and ethnic group as the outcome variable were evaluated. Most of the VIF values for the hypothesized models were below five (i.e., 1.01 to 2.22), indicating that multicollinearity is not an issue.

However, when total student enrollment and the number of male students were included then multicollinearity was present, thus, these two covariates were removed from the models. In terms of linearity, visual inspection of scatterplots did not indicate a nonlinear or curvilinear pattern. However, correlation coefficients between the modified BoQ-R total score and the average ODRs ranged from -0.01 to .10 indicating a weak relationship between the two variables. Notably, average ODRs from Asian students resulted in the only negative correlation coefficient (i.e., -0.01) indicating a weak negative relationship between PBIS implementation and the occurrence of exclusionary discipline. In terms of independence, visual inspection of scatterplots did not indicate a clear pattern or a cluster at certain points. Similarly, none of the results from the Durbin-Watson statistical test indicated the presence of autocorrelation in the residuals. In terms of homoscedasticity, visual inspection of scatterplots did indicate a pattern of variability based on the value of the modified BoQ-R total score for some groups. This is further supported by results from the Breusch-Pagan statistical test which identified residuals from Asian and Hispanic/Latine groups as having evidence of heteroscedasticity. In terms of the final assumption, normality, visual inspection of histograms and Q-Q plots did indicate the presence of non-normality. Specifically, the histograms did not appear fully bell-shaped but were skewed to the right. Additionally, the set of points on the Q-Q plot deviated from a straight line. Values from the skewness and kurtosis of the residuals for each model further support the presence of non-normality. Furthermore, the Shapiro-Wilk normality test was statistically significant for every model which means the residuals are considered to be non-normally distributed. To address the violations of the homoscedasticity and normality assumptions, removal of outliers and log-transformations of the outcome variable were both used for all models using (1) the modified BoQ-R total score as a predictor and (2) average ODRs per racial and ethnic group as the outcome variable.

With regard to the detection of outliers, results box plots were used to detect values for the outcome variable that fell outside the interquartile range using the cutoffs of 1.5 times below the 25th percentile and 1.5 times above the 75th percentile (Yang et al., 2019). For models using average ODRs as an outcome variable, the number of cases detected as outliers depended on the racial and ethnic groups. For Asian students, these models included seven outliers. For Black students, these models included four outliers. For Hispanic students, these models included five outliers. For Indigenous/Native students, these models included seven outliers. For multiracial students, these models included three outliers. For Pacific Islander students, these models included seven outliers. For White students, these models included five outliers. Since none of the detected outliers for each racial and ethnic group were greater than 5% of the data, then they were simply removed from analyses. The removal of these outliers allowed the homoscedasticity assumption to be met but non-normality and skewness of the data remained.

To address the violated normality assumption, log-transformations were conducted on all outcome variables. Using log-transformations is a common approach in multiple linear regression when values from the outcome variable are heavily skewed (Benoit, 2011). For the present analyses, log-transformations resulted in data that meet the normality assumptions. These transformations can complicate interpretation because it is different from interpreting regression coefficients that do not have log-transformed variables regression coefficients. Specifically, the unstandardized regression coefficients in the present study are interpreted as the percent increase (or decrease) in the response for every one-unit increase in the independent variable.

Table 2.2 includes the unstandardized regression coefficients for each model describing the relationship between overall PBIS implementation fidelity, relevant covariates, and the occurrence of exclusionary discipline. There were no statistically significant estimates amongst these covariates or the focal predictor in models evaluating the average number of ODRs for Black, Pacific Islander, and White students. Models from the remaining racial and ethnic groups resulted in at least one statistically significant predictor. However, it is important to note that the adjusted R-squared values ranged from .016 to .100 which means that the final models minimally explained the variance of average ODRs per racial and ethnic group.

For Asian students, the modified total BoQ-R score (b = -0.00283, SE = 0.00121, p = .020) was significantly and negatively related to the log-transformed average number of ODRs received by Asian students. Thus, on average, for each one-unit increase in the modified total BoQ-R score, average Asian

ODRs decreased by 0.28 percent. None of the covariates in these models for Asian students were statistically significant.

For Hispanic/Latine students, the covariates of Hispanic/Latine student enrollment (b = -0.00044, SE = 0.00018, p = .016) were both significantly and negatively related to log-transformed average ODRs received by these students. Thus, on average, for each one-unit increase in the number of Hispanic/Latine students enrolled, average Hispanic/Latine ODRs decreased by 0.04 percent. Also, the covariate of high school (b = 0.31335, SE = 0.14254, p = .033) was significantly and positively related to log-transformed average ODRs received by Hispanic/Latine students. Thus, on average, compared to Hispanic/Latine students enrolled in elementary schools, Hispanic/Latine students enrolled in high schools received a 36.80 percent increase in average ODRs. None of the remaining covariates significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the occurrence of average ODRs received by Hispanic/Latine students.

For Indigenous/Native students, the covariate of middle school (b = 0.29885, SE = 0.14088, p =.049) was significantly and positively related to log-transformed average ODRs received by Indigenous/Native students. Thus, on average, compared to Indigenous/Native students enrolled in elementary schools, Indigenous/Native students enrolled in middle schools received a 34.83 percent increase in average ODRs. None of the remaining covariates significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the occurrence of average ODRs received by Indigenous/Native students.

For multiracial students, the covariate of middle school (b = 0.22900, SE = 0.08655, p = .01) significantly and positively related to log-transformed average ODRs received by these students. Thus, on average, compared to multiracial students enrolled in elementary schools, multiracial students enrolled in middle schools received a 25.73 percent increase in average ODRs. None of the remaining covariates significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the occurrence of average ODRs received by multiracial students.

Research Question 2: PBIS Implementation Components and Occurrence of Exclusionary Discipline

Assumptions for models using (1) the total scores from each BoQ-R critical element as predictors and (2) average ODRs per racial and ethnic group as the outcome variable were evaluated. Most of the VIF values for the hypothesized models were below five (i.e., 1.06 to 3.30). However, when total student enrollment and the number of male students were included then multicollinearity was present, thus, these two covariates were removed from the models. In terms of linearity, visual inspection of scatterplots did not indicate a nonlinear or curvilinear pattern. However, correlation coefficients between total scores on the BoQ-R critical elements and the average ODRs ranged from 0.00 to 0.12 indicating a weak positive relationship between the two variables. Notably, none of the correlation coefficients for average ODRs from Black, Indigenous/Native or White demonstrated any negative relationships between implementation of specific PBIS components and the occurrence of exclusionary discipline. In terms of independence, visual inspection of scatterplots did not indicate a clear pattern or a cluster at certain points. Similarly, none of the results from the Durbin-Watson statistical test indicated the presence of autocorrelation in the residuals. In terms of homoscedasticity, visual inspection of scatterplots did not indicate a pattern of variability based on the values of the BoQ-R critical elements. However, results from the Breusch-Pagan statistical test identified residuals from the Hispanic/Latine group as having evidence of heteroscedasticity. In terms of the final assumption, normality, visual inspection of histograms and Q-Q plots did indicate the presence of non-normality. Specifically, the histograms did not appear fully bell-shaped but were skewed to the right. Also, the set of points on the Q-Q plot deviated from a straight line. Values from the skewness and kurtosis of the residuals for each model further support the presence of non-normality. Furthermore, the Shapiro-Wilk normality test was statistically significant for every model which means the residuals are considered to be non-normally distributed. To address the violations of the homoscedasticity and normality assumptions, removal of outliers and log-transformations of the outcome variable were both used for all models using 1) the total scores from each BoQ-R critical element as predictors and (2) average ODRs per racial and ethnic group as the outcome variable. The process of outlier removals and log-transformations of the outcome variable were described earlier.

Table 2.3 includes the unstandardized regression coefficients for each model describing the relationship between implementation of specific PBIS outcomes, relevant covariates, and the occurrence of exclusionary discipline. There were no statistically significant estimates amongst these covariates or the focal predictors in models evaluating the average number of ODRs for Black students. Models from the remaining racial and ethnic groups resulted in at least one statistically significant predictor. However, it is important to note that the adjusted R-squared values ranged from .025 to .120 which means that the final models minimally explained the variance of average ODRs per racial and ethnic group.

For Asian students, the BoQ-R critical system of *Data Entry* (b = 0.03452, SE = 0.01616, p =.034) significantly and positively related to log-transformed average ODRs received by Asian students. Thus, on average, for each one-unit increase in the total score of *Data Entry*, average Asian ODRs increased by 3.51 percent. None of the covariates, or remaining BoQ-R critical systems, significantly related to the occurrence of average Asian ODRs.

For Hispanic/Latine students, the covariate of high school (b = 0.24060, SE = 0.11757, p = .043) significantly and positively related to log-transformed average ODRs received by Hispanic/Latine students. Thus, on average, compared to Hispanic/Latine students enrolled in elementary schools, Hispanic/Latine students enrolled in high schools received a 27.20 percent increase in average ODRs. None of the remaining covariates, nor any of the BoQ-R critical systems, significantly related to the occurrence of average Hispanic/Latine ODRs.

For Indigenous/Native students, the covariate of middle school (b = 0.24585, SE = 0.11956, p = .042) significantly and positively related to log-transformed average ODRs received by Indigenous/Native students. Thus, on average, compared to Indigenous/Native students enrolled in elementary schools, Indigenous/Native students enrolled in middle schools received a 27.87 percent increase in average ODRs. None of the remaining covariates, nor any of the BoQ-R critical systems, significantly related to the occurrence of average Indigenous/Native ODRs.

For multiracial students, the covariate of middle school (b = 0.24070, SE = 0.08798, p = .007) significantly and positively related to log-transformed average ODRs received by multiracial students. Thus, on average, compared to multiracial students enrolled in elementary schools, multiracial students enrolled in middle schools received a 27.21 percent increase in average ODRs. None of the remaining covariates, nor any of the BoQ-R critical systems, significantly related to the occurrence of average multiracial ODRs.

For Pacific Islander students, the covariate of high school (b = 0.27827, SE = 0.13371, p = .040) significantly and positively related to log-transformed average ODRs received by Pacific Islander students. Thus, on average, compared to Pacific Islander students enrolled in elementary schools, Pacific Islander students enrolled in high schools received a 32.08 percent increase in average ODRs. This was the only covariate that significantly related to the occurrence of average Pacific Islander ODRs. The BoQ-R critical system of *Expectations and Rules* (b = -0.08907, SE = 0.04020, p =.029) significantly and negatively related to log-transformed average ODRs received by Pacific Islander students. Thus, on average, for every one-unit increase in total score of *Expectations and Rules*, average Pacific Islander ODRs decreased by 8.52 percent. None of the remaining BoQ-R critical systems significantly related to the occurrence of average Pacific Islander ODRs.

For White students, the BoQ-R critical system of *Expectations and Rules* (b = -0.08660, SE = 0.03979, p =.031) significantly and negatively related to log-transformed average ODRs received by White students. Thus, on average, for every one-unit increase in total score of *Expectations and Rules*, average White ODRs decreased by 8.30 percent. None of the remaining BoQ-R critical systems, nor any of the covariates, significantly related to the occurrence of average White ODRs.

Research Question 3: Overall PBIS Implementation and (In)equitable Discipline Outcomes

Assumptions for models using (1) the modified BoQ-R total score as a predictor and (2) risk for ODRs per racial and ethnic group as the outcome variable were evaluated. All of the VIF values for the hypothesized models were below five (i.e., 1.01 to 2.23). However, when total student enrollment and the number of male students were included then multicollinearity was present, thus, these two covariates were removed from the models. In terms of linearity, visual inspection of scatterplots did not indicate a nonlinear or curvilinear pattern. However, correlation coefficients between the modified BoQ-R total score and the risk for ODRs ranged from -0.03 to -0.19 indicating a weak negative relationship between the two variables. Notably, ODR risk for Asian students resulted in the only positive correlation coefficient (i.e., 0.06) indicating a weak positive relationship between PBIS implementation and the occurrence of exclusionary discipline. In terms of independence, visual inspection of scatterplots did not indicate a clear pattern or a cluster at certain points. Similarly, none of the results from the Durbin-Watson statistical test indicated the presence of autocorrelation in the residuals. In terms of homoscedasticity, visual inspection of scatterplots did indicate a pattern of variability based on the value of the modified BoQ-R total score for some groups. This is further supported by results from the Breusch-Pagan statistical test which identified residuals from Asian, Black, and Hispanic/Latine groups as having evidence of heteroscedasticity. In terms of the final assumption, normality, visual inspection of histograms and Q-Q plots did indicate the presence of non-normality. Specifically, the histograms did not appear fully bell-shaped but were skewed to the right. Also, the set of points on the Q-Q plot deviated from a straight line. Values from the skewness and kurtosis of the residuals for each model further support the presence of non-normality. Furthermore, the Shapiro-Wilk normality test was statistically significant for every model which means the residuals are considered to be non-normally distributed. To address the violations of the homoscedasticity and normality assumptions, removal of outliers and log-transformations of the outcome variable were both used for all models using (1) the modified BoQ-R total score as a predictor and (2) risk for ODRs per racial and ethnic group as the outcome variable.

For models using risk ratios for ODRs as an outcome variable (with White students as the reference group), the number of cases detected as outliers depended on the racial and ethnic groups. For Asian students, these models included two outliers. For Black students, these models included two outliers. For Hispanic students, these models included five outliers. For Indigenous/Native students, these models included four outliers. For multiracial students, these models included two outliers. For Pacific Islander students, these models included 15 outliers. Again, none of the detected outliers for each racial and ethnic group were greater than 5% of the data, so they were simply removed from analyses. The removal of these outliers allowed the homoscedasticity assumption to be met but non-normality and skewness of the data remained.

Table 2.4 includes the unstandardized regression coefficients for each model describing the relationship between overall PBIS implementation fidelity, relevant covariates, and the risk for exclusionary discipline. There were no statistically significant estimates amongst these covariates or the predictor in models evaluating the risk ratios for multiracial students receiving ODRs. Models from the remaining racial and ethnic groups resulted in at least one statistically significant predictor. However, it is important to note that the adjusted R-squared values ranged from .025 to .118 which means that the final models minimally explained the variance of risk for ODRs per racial and ethnic group.

For Asian students, the covariate of Asian student enrollment (b = 0.00092, SE = 0.00045, p = .043) significantly and positively related to the log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the number of Asian students enrolled, the ODR risk for Asian students increased by .09 percent. However, the modified total BoQ-R score (b = -0.00340, SE = 0.00129, p = .009) significantly and negatively related to the log-transformed risk of Asian students receiving ODRs. Thus, on average, for each one-unit increase in the modified total BoQ-R score, the ODR risk for Asian students decreased by .34 percent. None of the remaining covariates in this model significantly related to the risk of Asian students receiving ODRs.

For Black students, the covariate of Black student enrollment (b = 0.00068, SE = 0.00033, p = .038) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the number of Black students enrolled, the ODR risk for Black students increased by .07 percent. However, the covariate of FRL (b = -0.42183, SE = 0.16375, p = .012) significantly and negatively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the percentage of FRL, the ODR risk for Black students decreased by 34.42 percent. None of the remaining covariates significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the risk of Black students receiving ODRs.

For Hispanic/Latine students, the covariate of high school (b = 0.20574, SE = 0.09326, p = .028) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, compared to Hispanic/Latine students enrolled in elementary schools, Hispanic/Latine students enrolled in high schools received a 22.84 percent increase in risk for receiving ODRs. This was the only covariate that significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the risk of Hispanic/Latine students receiving ODRs.

For Indigenous/Native students, the covariate of high school (b = 0.48373, SE = 0.15500, p = .002) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, compared to Indigenous/Native students enrolled in elementary schools, Indigenous/Native students enrolled in high schools received a 62.21 percent increase in risk for receiving ODRs. This was the only covariate that significantly related to the outcome variable. Our focal predictor in this model, the modified total BoQ-R score, did not significantly relate to the risk of Indigenous/Native students receiving ODRs.

For Pacific Islander students, the covariate of high school (b = 0.34203, SE = 0.12461, p = .007) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, compared to Pacific Islander students enrolled in elementary schools, Pacific Islander students enrolled in high schools received a 40.78 percent increase in risk for receiving ODRs. However, the modified total BoQ-R score (b = -0.00495, SE = 0.00215, p = .023) significantly and negatively related to the log-transformed risk of Pacific Islander students receiving ODRs. Thus, on average, for each one-unit increase in the modified total BoQ-R score, the ODR risk for Pacific Islander students decreased by .49 percent. None of the remaining covariates in this model significantly related to the risk of Pacific Islander students receiving ODRs.

Research Question 3: PBIS Implementation Components and (In)equitable Discipline Outcomes

Assumptions for models using (1) the total scores from each BoQ-R critical element as predictors and (2) risk for ODRs per racial and ethnic group as the outcome variable were evaluated. Most of the VIF values for the hypothesized models were below five (i.e., 1.06 to 3.30). However, when total student enrollment and the number of male students were included then multicollinearity was present, thus, these two covariates were removed from the models. In terms of linearity, visual inspection of scatterplots did not indicate a nonlinear or curvilinear pattern. However, correlation coefficients between total scores on the BoQ-R critical elements and risk for ODRs ranged from -0.00 to -0.19 indicating a weak negative relationship between the two variables. Notably, ODR risk for Indigenous/Native, Multiracial, and Pacific Islander students resulted in some positive correlation coefficients, indicating a positive relationship between implementation of specific PBIS components and the occurrence of exclusionary discipline. In terms of independence, visual inspection of scatterplots did not indicate a clear pattern or a cluster at certain points. Similarly, none of the results from the Durbin-Watson statistical test indicated the presence of autocorrelation in the residuals. In terms of homoscedasticity, visual inspection of scatterplots did not indicate a pattern of variability based on the values of the BoQ-R critical elements. However, results from the Breusch-Pagan statistical test identified residuals from the Black group as having evidence of heteroscedasticity. In terms of the final assumption, normality, visual inspection of histograms and Q-Q plots did indicate the presence of non-normality. Specifically, the histograms did not appear fully bell-shaped but skewed to the right. Also, the set of points on the Q-Q plot deviated from a straight line. Values from the skewness and kurtosis of the residuals for each model further support the presence of non-normality. Furthermore, the Shapiro-Wilk normality test was statistically significant for every model which means the residuals are considered to be non-normally distributed. To address the violations of the homoscedasticity and normality assumptions, removal of outliers and log-transformations of the outcome variable were both used for all models using (1) the total scores from each BoQ-R critical element as predictors and (2) risk for ODRs per racial and ethnic group as the outcome variable.

Table 2.5 includes the unstandardized regression coefficients for each model describing the relationship between implementation of specific PBIS outcomes, relevant covariates, and the risk for exclusionary discipline. There were no statistically significant estimates amongst these covariates or the predictor in models evaluating the risk ratios for Asian students receiving ODRs. Models from the remaining racial and ethnic groups resulted in at least one statistically significant predictor. However, it is important to note that the adjusted R-squared values ranged from .024 to .223 which means that the final models minimally explained the variance of risk for ODRs per racial and ethnic group.

For Black students, the covariate of Black student enrollment (*b* = 0.00069, *SE* = 0.00034, *p* = .046) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the number of Black students, the ODR risk for Black students increased by .07 percent. However, the covariate of FRL (*b* = -0.34271, *SE* = 0.15537, *p* = .029) significantly and negatively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the percentage of students receiving FRL, the ODR risk for Black students decreased by 29.02 percent. None of the remaining covariates, nor any of the BoQ-R critical systems, significantly related to the risk of Black students receiving ODRs.

For Hispanic/Latine students, the covariate FRL (*b* = -0.28385, *SE* = 0.13532, *p* = .038) significantly and negatively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit in the percentage of students receiving FRL, the ODR risk for Hispanic/Latine students decreased by 24.71 percent. None of the remaining covariates, nor any of the BoQ-R critical systems, significantly related to the risk of Hispanic/Latine students receiving ODRs.

For Indigenous/Native students, the covariate of high school (*b* = 0.41019, *SE* = 0.16400, *p* = .014) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, compared to Indigenous/Native students enrolled in elementary schools, Indigenous/Native students enrolled in high schools received a 50.71 percent increase in risk for receiving ODRs. However, the BoQ-R critical system of *Lesson Plans* (*b* = -0.07885, *SE* = 0.03241, *p* = .017) significantly and negatively related to log-transformed risk of Indigenous/Native students receiving ODRs. Thus, on average, for each one-unit increase in the total score of *Lesson Plans*, the ODR risk for Indigenous/Native students decreased by 7.58 percent. None of the remaining BoQ-R critical systems, nor covariates, significantly relate to the risk of Indigenous/Native students receiving ODRs.

For multiracial students, the BoQ-R critical system of *Effective Discipline Procedures* (*b* = -0.06287, *SE* = 0.02982, *p* = .036) significantly and negatively related to log-transformed risk of these students receiving ODRs. Thus, on average, for each one-unit increase in the total score of *Effective Discipline Procedures*, the ODR risk for multiracial students decreased by 6.09 percent. None of the remaining BoQ-R critical systems, nor any covariates, significantly relate to the risk of multiracial students receiving ODRs.

For Pacific Islander students, the covariate of high school (b = 0.39279, SE = 0.12595, p = .003) significantly and positively related to log-transformed risk of these students receiving ODRs. Thus, on average, compared to Pacific Islander students enrolled in elementary schools, Pacific Islander students enrolled in high schools received a 48.11 percent increase in risk for receiving ODRs. Also, the BoQ-R critical system of *Effective Discipline Procedures* (b = 0.07359, SE = 0.03433, p = .035) significantly and positively related to log-transformed risk of Pacific Islander students receiving ODRs. Thus, on average, for each one-unit increase in the total score of *Effective Discipline Procedures*, the ODR risk for Pacific Islander students increased by 7.64 percent. However, the BoQ-R critical system of *Expectations and Rules* (b = -0.10252, SE = 0.03392, p = .003) significantly and negatively related to log-transformed risk of Pacific Islander students receiving ODRs. Thus, on average, for each one-unit increase in the total score of *Expectations and Rules*, the ODR risk for Pacific Islander students decreased by 9.74 percent. None of the remaining BoQ-R critical systems, nor covariates, significantly relate to the risk of Pacific Islander students receiving ODRs.

Discussion

Despite PBIS being an implementation framework commonly used in schools across the U.S., the different components of PBIS being implemented can vary greatly. This is why it is important for school professionals to monitor the fidelity of PBIS implementation. Fortunately, there are many PBIS implementation fidelity measures for school professionals to use, including the BoQ (Cohen et al., 2007). The purpose of the BoQ is to evaluate implementation of Tier 1 PBIS components. Important revisions were made to this measure in 2011 (i.e., BoQ-R) which included replacing the *Crisis Plan* critical element with the *Classroom Systems* critical element (Childs et al., 2011). The present study (1) evaluated the psychometric evidence of the BoQ-R using findings from a national sample of schools implementing PBIS and (2) explored the association between overall scores as well as critical elements from the BoQ-R and exclusionary discipline.

Psychometric Evidence of the BoQ-R

Findings from the present study initially provided little support for the hypothesized factor structure of the BoQ-R as presented by the developers of this measure (Childs et al., 2011). However, once a single item was removed (i.e., item 8) then there was compelling evidence for the structural validity of the BoQ-R in a national sample. The need to remove item 8 aligns with CFA findings from Barclay (2017) which reported this item as having the only negative factor loading. After removing this item, the remaining 52 items aligned with the 10 critical elements and sufficiently captured the variance of data collected in a national sampling of schools implementing PBIS. This is the first study to the researchers' knowledge to evaluate the factor structure of the BoQ-R with previous studies relying on findings from the initial BoQ measure (Cohen et al., 2007) when reporting psychometric properties. That being said, additional research on the factor structure of the BoQ-R with schools implementing PBIS, but not using SWIS, are needed in light of the present study's findings. Also, while researchers originally planned to include all 10 subscales from the BoQ-R into the final regression models, the psychometric evidence did not support the inclusion of scores from the *PBIS Team* and *Faculty Commitment* subscales. Additional research into why the items from these subscales resulted in low reliability is needed. Ultimately, the psychometric evidence in the present study supported use of the modified total BoQ-R score as well as eight out of the 10 subscales. These were the scores included in the final models.

The BoQ-R and Exclusionary Discipline

The relation between the BoQ-R and exclusionary discipline varied across racial and ethnic groups even when taking into account the number of students of color enrolled, FRL rates, and grade-level. For example, when the modified total BoQ-R score increased, then the average number of ODRs received by Asian students decreased. When total scores on a specific BoQ-R critical element, *Expectations and Rules*, increased then the average number of ODRs received by Pacific Islander and White students decreased. However, this finding was not replicated with any of the remaining racial and ethnic groups. Plus, descriptive analyses illustrated that Pacific Islander students already received the fewest ODRs compared to all other groups. Still, this finding that higher implementation fidelity in the *Expectations and Rules* critical element then resulted in decreased occurrence of ODRs for students from certain racial and ethnic groups, is consistent with previous finings that teaching school-wide rules and expectations for behavior is an 'active ingredient' for PBIS implementation (Molloy et al., 2013).

In terms of equitable discipline practices, Asian and Pacific Islander were the only racial and ethnic groups that yielded significant findings between the modified total BoQ-R score and ODR risk. Specifically, when the modified total BoQ-R score increased, then the risk for Asian and Pacific Islander students receiving ODRs compared to White students decreased. It is important to note that prior to these analyses, results from descriptive analyses already showed Asian students to be at the same risk of receiving an ODR compared to White students and Pacific Islander students to be at a lower risk of receiving an ODR compared to White students. Thus, these findings do not support that overall PBIS implementation improved racial discipline disparities in the occurrence of exclusionary discipline. This is consistent with previous findings that race-neutral or race-evasive approaches should not be expected to automatically yield equitable outcomes (Cruz & Firestone, 2023; Skiba, 2015). The reliance on such approaches dominates the current perspectives and approaches taken in education and disparities remain.

The pattern in findings between the specific components of PBIS and ODR risk were more promising. When total scores on the BoQ-R critical element, *Lesson Plans*, increased then the risk for Indigenous/Native students receiving ODRs compared to White students decreased. Plus, prior to these analyses, results from descriptive analyses already showed Indigenous/Native students to be at a higher risk of receiving an ODR compared to White students. So, this finding illustrates that greater implementation fidelity in the PBIS component, *Lesson Plans*, resulted in greater equity in exclusionary discipline for Indigenous/Native students. This pattern continued when total scores on the modified BoQ-R critical element, *Effective Discipline Procedures*, increased then the risk for multiracial students receiving ODRs compared to White students decreased. Plus, prior to these analyses, results from descriptive analyses already showed multiracial students to be at a higher risk of receiving an ODR compared to White students. So, this finding illustrates that greater implementation fidelity in the PBIS component, *Effective Discipline Procedures* (without item 8), resulted in greater equity in exclusionary discipline for multiracial students. However, this finding was not replicated with Pacific Islander students. These students experienced greater inequity in exclusionary discipline when total scores on the modified BoQ-R critical element, *Effective Discipline Procedures*, increased because then the risk for this group of students receiving ODRs compared to White students also increased. Again, descriptive analyses demonstrated that Pacific Islander students tended to be at the same risk for receiving ODRs as White students. Yet, when total scores on the BoQ-R critical element, *Expectations and Rules*, increased then this resulted in greater equity in exclusionary discipline for Pacific Islander students.

Overall, there is evidence that specific PBIS practices from the *Lesson Plans*, the modified *Effective Discipline Procedures*, and the *Expectations and Rules* critical elements related to improved equity for Indigenous/Native, multiracial students, and Pacific Islander students respectively. However, there is conflicting evidence regarding higher implementation of the *Effective Discipline Procedures* critical element because it resulted in greater inequity for Pacific Islander students. The lack of previous research supporting these findings for these specific groups, underscores the importance of future studies disaggregating racial and ethnic groups as much as possible instead of using an aggregate variable (i.e., 'non-White').

The results in this study did not replicate findings in Barclay et al. (2022) related to the BoQ-R critical element, *Classroom Systems*, significantly relating to any of the outcome variables across racial and ethnic groups. This lack of replication could be a result of using log-transformed outcome variables to account for the non-normality in the present study's data. Also, some of the demographic characteristics that describe the present study's data may explain the violation of the normality assumption. For example, all participating schools included in the dataset submitted a complete BoQ-R survey and agreed to have their results used for research purposes. This might have resulted in selection bias that led to researchers needing to account for non-normality in the regression models. The present study's sample also included schools that were predominantly White. The lack of racial and ethnic diversity within participating schools could also explain the minimal R-squared values in the present study. Notably, this is not the first study to use ODRs as the primary outcome variable and have models that inadequately explain variation based on data. Specifically, Barclay et al. (2022) stated that the "amount of variance accounted for was not significant for models of ODR risk ratios" (p. 8). Prior studies tend to focus on the occurrence of suspensions and expulsions, so the extent to which ODRs fully capture racial discipline disparities and other phenomenon related to the discipline gap is an area that need to be further studied.

Limitations

It is important to note some limitations of the present study. First it is important to consider features of the analytic sample. The composition of schools included in the analysis all used SWIS, which is not representative of all schools. It is likely that schools that allocated funds to use SWIS differ systematically in their implementation of PBIS than schools that do not. Furthermore, all data were reported at the school-level, which made it impossible to interpret findings for individual students. Thus, all findings must be interpreted under the context of the impact of scores from the BoQ-R on specific racial and ethnic groups across schools not students. Also, not all schools reported student enrollment for specific racial and ethnic groups that led to a significant amount of missing data. The missing data in the outcome variables is a major limitation. However, researchers attempted to address this limitation via use of multiple imputation. Since all of the participating schools came from a public school district, then the results from the present study are not generalizable to private schools. Additionally, majority of schools in the present study came from a suburban geographical location with a limited number of schools located in urban areas. The participating schools also reported a higher-than-average score on the BoQ-R (i.e., greater than 70%; Kincaid & George, 2010) indicating that schools in the present study experienced high implementation fidelity of Tier 1 PBIS. Again, this suggests systematic differences with regard to PBIS implementation of schools included in the dataset.

It is also important to consider limitations of the study design and methods. Unfortunately, the cross-sectional design of the present study is a limitation that prevented causal interpretation and prevented researchers from understanding how response to implementation of PBIS over time was associated with exclusionary discipline. The present study did not account for multiple comparisons made, given the paucity of research on the BoQ-R subscales, and as a result inflated Type 1 error likely occurred. Researchers decided not to account for multiple comparisons as few findings would then be considered ‘statistically significant’ and results from the present study are exploratory and preliminary. Future research is greatly needed exploring the BoQ-R given findings of the present study. Finally, when measuring (in)equity in discipline, the present studied relied on risk ratios with White students as the comparison group. Opinions on which group(s) to use as the comparison group for risk ratios greatly vary. However, given the influence of White supremacy in school systems and the inequitable use of exclusionary discipline practices towards Black students in particular, the reference group in the present study was chosen accordingly. In light of these limitations, it is important for future research to continue investigating under what conditions implementation of PBIS can address the persistent racial discipline disparities in schools.

Future Directions

Research

Researchers should continue to evaluate the association between implementation of PBIS and exclusionary discipline. While results from the present study provides some support for this relation, future research should further investigate an emerging consistent finding in the literature that ODRs as an outcome variable tend to lead to insignificant results compared to suspensions as an outcome variable for exclusionary discipline (Barclay et al., 2022). It is also important for future studies to continue to provide psychometric evidence for the BoQ-R especially since this measure was revised from when the original psychometric evidence was evaluated (Childs et al., 2011).

Practice

If practitioners implementing PBIS in schools uses the BoQ-R to monitor their implementation fidelity, then they should consider potential shortcomings of the measure, including this study’s finding that some subscales have low internal consistency. This is especially the case for schools utilizing SWIS to document their discipline procedures. Despite the psychometric evidence supporting its use, information from the overall BoQ-R score as well as scores from the critical systems inconsistently related to the occurrence of exclusionary discipline for certain racial and ethnic groups in the present study. This suggests that practitioners should not expect the implementation of PBIS in and of itself to mitigate racial discipline disparities that exist in their school system, especially for students across various racial and ethnic groups.

Conclusions

Findings from the present study supported an association between implementation of PBIS and exclusionary discipline for certain racial and ethnic groups. Similar to previous findings in the literature that overall PBIS implementation does little to impact racial discipline disparities, our results support that the same is true for the implementation of certain components or critical elements of PBIS. Despite evidence for the psychometric properties of the BoQ-R in the present study's national sample, there lacked consistent significant findings across racial and ethnic groups. These results align with previous studies that also used ODRs as the primary outcome variable for exclusionary discipline instead of suspensions or expulsions. Unfortunately, how well schools implemented PBIS did not lead to a greater reduction or improved equity in their exclusionary discipline data. Overall, it is essential to ensure that the systems, policies, and practices in schools are explicitly centered around equity if one wants to experience equitable outcomes.

Chapter 4

Synthesis and General Discussion

Studies 1 and 2 do not provide evidence to suggest that the classroom behavior support practices considered can mitigate racial discipline disparities. Traditionally, researchers and practitioners have prioritized implementing race-neutral or race-evasive school-wide interventions (Bradshaw et al., 2012; Flannery et al., 2014; Molloy et al., 2013; Noltmeyer et al., 2019) to improve student behavior (Childs et al., 2016; Gregory et al., 2016; Mitchell & Bradshaw, 2013; Pane et al., 2015). The empty review from study 1 and the lack of consistent significant estimates from study 2 illuminate the need for researchers and practitioners to prioritize evaluating how implementation of classroom behavior management interventions influence not only student behavior, but teacher behavior. This fundamental shift is a necessary one to begin understanding how racism, ableism, and privilege perpetuate inequities in schools. Findings from both studies add to the growing evidence-base that implementing procedures that were not intentionally design to center equity and decenter oppression are insufficient in yielding equitable outcomes. It is critical for both school-based professionals and researchers to begin thinking critically and creatively about how to leverage one's understanding of systemic oppression to effectively address the persistent racial discipline disparities in schools.

Implications for Research

Both studies attempted to fill in gaps in the current research literature by (1) identifying the need for exclusionary discipline measures to be used as outcome variables in classroom behavior management research and (2) identifying the need to disaggregate racial and ethnic groups when evaluating racial discipline disparities. In addition to findings from study 1 highlighting that researchers did not report exclusionary discipline measures as outcome variables in the context of the CW-FIT, findings from study 2 provided further evidence that whether the use of risk ratios for office discipline referrals (ODRs) were significantly impacted by school implementation of positive behavioral interventions and supports (PBIS) depended on the specific racial and ethnic group. Previous studies often focus on a couple of racial and ethnic groups (mainly Black and Hispanic/Latine) despite their being empirical evidence that racial discipline disparities exist beyond a couple of racial and ethnic groups. The overall findings from both studies further underscore the importance of researchers comprehensively evaluating racial discipline disparities by disaggregating and including results from as many racial and ethnic groups as possible. Results from the first study highlight problems associated with researchers under-reporting information (i.e., impact of the CW-FIT on use of exclusionary discipline practices) so results from the second study attempted to avoid this problem by reporting information on as many racial and ethnic groups as possible (i.e., Asian, Black, Hispanic/Latine, Indigenous/Native, multiracial, Pacific Islander, White).

Implications for Practice

Both studies focused on interventions (i.e., CW-FIT) and approaches (i.e., PBIS) for school-based professionals to use in their practice to effectively manage student behavior. Yet, neither results from implementing the CW-FIT nor PBIS supported effectively reducing the use of exclusionary discipline practices and/or addressing the disparate use of these practices in schools. Thus, it is critical for practitioners to refrain from assuming that the implementation of classroom or school-wide behavior management practices automatically address the racial discipline disparities that have persisted in U.S. schools for decades. Instead, to improve the likelihood that there will be improved equity in discipline outcomes, those in practice should consider implementing interventions and approaches that are explicitly culturally responsive and equity-focused (Muldrew & Miller, 2021; Sullivan et al., 2020).

Future Research Directions

The continued gaps in the literature should be addressed by future researchers on how to leverage classroom behavior management interventions to improve equitable discipline outcomes. Findings from both studies illustrate that it is necessary for future researchers to prioritize outcome variables that could prevent the occurrence of students experiencing more severe forms of exclusionary discipline (i.e., ODRs). Findings from study 1 support the need for future research to directly monitor the association between teacher-implemented classroom behavior management and the occurrence of exclusionary discipline. Findings from study 2 support the need for future research to ensure data are evaluated amongst various racial and ethnic groups.

Conclusion

It is well-documented that teachers are sorely underprepared to manage the disruptive student behavior that occurs in their classroom. It is also well-documented that the use of exclusionary discipline, by teachers, are inequitable across racial and ethnic groups. Instead of focusing on implementing behavior management interventions (e.g., CW-FIT, PBIS) to improve student behavior, more research is needed to evaluate the association between implementing these types of interventions in improving teacher behavior. Hopefully, this shift towards encouraging teacher behavior change, as opposed to a sole focus on demanding child behavior change, can more effectively bring us closer to closing the discipline gap.

Table 1.1

Descriptive Characteristics of Included Studies

Authors & Publication Year	Non-White Student Participants (%)	Grade-level(s)	Locale	Design	Study Quality Score (%)*
1. Caldarella et al. (2015)	Not Reported	LE	Suburban	GD	100%
2. Caldarella et al. (2017)	77%	M	Not Reported	SCRD	95%
3. Conklin et al. (2017) †	Not Reported	LE + M	Urban	SCRD	95%
4. Hansen et al. (2017) †	Not Reported	LE + UE	Not Reported	SCRD	95%
5. Hirsch et al. (2016)	Not Reported	LE	Urban	SCRD	100%
6. Kamps et al. (2011) †	54%	LE + UE	Urban + Suburban	SCRD	91%
7. Kamps et al. (2015a)	89%	LE + UE	Urban	SCRD	95%
8. Kamps et al. (2015b)	Not Reported	LE + UE	Urban	GD	88%
9. Monson et al. (2020)	37%	M	Not Reported	SCRD	100%
10. Naylor et al. (2018)	96%	LE	Urban	SCRD	95%
11. Nelson et al. (2018) †	62%	UE	Suburban	SCRD	91%
12. Orr et al. (2020)	35%	M + H	Suburban	SCRD	100%
13. Parikh (2019)[a†]	Not Reported	E	Not Reported	SCRD	91%
14. Speight et al. (2021)	36%	H	Not Reported	SCRD	100%
15. Speight et al. (2020) †	76%	M	Not Reported	SCRD	100%
16. Weeden et al. (2016)	50%	LE + UE	Urban	SCRD	100%
17. Wills et al. (2014) †	28%	LE	Urban	SCRD	95%
18. Wills et al. (2018)	55%	LE + UE + M	Not Reported	GD	96%
19. Wills et al. (2022)	Not Reported	LE + UE	Not Reported	SCRD	82%

Note. †Study included in Quantitative Synthesis; [a]Master's thesis; LE = Lower Elementary; UE = Upper Elementary; M = Middle; H = High school; LE is grades K-2; UE is grades 3-5; and M is grades 6-8; *Percentage calculated out of 22 OR 24 based on Quality Indicators by CEC Standards; GD (group design) = Total of 24 potential points; SCRD (single-case research design) = Total of 22 potential points

Table 1.2

Effectiveness of Tier-1 CW-FIT implementation in classrooms within SCRD studies

Study	*M* (SD)		Tau/Tau-U		PAND	
	Baseline	Intervention	Estimate	Effectiveness	Estimate	Effectiveness
Caldarella et al. (2017)	*55.43* (11.75)	*80.90* (8.9)	0.91	75th Percentile	0.95	90th Percentile
Conklin et al. (2017)	*41.18* (18.28)	*86.43* (9.8)	1.00	90th Percentile	1.00	90th Percentile
Hansen et al. (20117)	*51.10* (14.9)	*70.97* (10.4)	0.87	75th Percentile	0.89	75th Percentile
Hirsch et al. (2016)	*34.37* (11.4)	*72.74* (13.0)	0.99	90th Percentile	0.96	90th Percentile
Kamps et al. (2011)	*46.06* (19.1)	*81.70* (10.9)	0.96	90th Percentile	0.96	90th Percentile
Kamps et al. (2015a)	*48.19* (14.9)	*82.83* (10.9)	0.95	90th Percentile	0.96	90th Percentile
Monson et al. (2020)	56.27 (16.16)	*84.90* (9.12)	0.92	75th Percentile	0.91	75th Percentile
Naylor et al. (2018)	*62.53* (14.2)	*87.39* (11.2)	0.82	75th Percentile	0.89	75th Percentile
Nelson et al. (2018)	*66.21* (12.8)	88.69 (6.5)	0.98	90th Percentile	0.96	90th Percentile
Orr et al. (2020)	65.43 (5.4)	78.18 (6.7)	0.79	75th Percentile	0.92	90th Percentile
Parikh (2019)	*63.56* (24.8)	*34.30* (23.2)	0.84*	75th Percentile	0.78	75th Percentile
Speight (2020)	*47.00* (11.8)	*86.93* (9.1)	1.00	90th Percentile	0.99	90th Percentile
Speight (2021)	57.08 (6.3)	82.51 (4.3)	1.00	90th Percentile	1.00	90th Percentile
Weeden et al. (2016)	*57.33* (13.4)	*90.86* (4.9)	0.98	90th Percentile	0.96	75th Percentile
Wills et al. (2014)	*63.74* (7.2)	*92.93* (5.6)	1.00	90th Percentile	1.00	90th Percentile
Wills et al. (2022)^	46.80^	77.04^	0.87*	75th Percentile	---	---
Hedge's *g* ES estimate:			*g* = 1.90			

Note. *Tau-U; ^Unable to independently calculate the mean, standard deviation, and effect sizes because the original authors did not provide time series graphs; ^Reported as an "overall average across classrooms" (Wills et al., 2022, p. 200); †Studies included in meta-analysis to calculate

Table 1.3

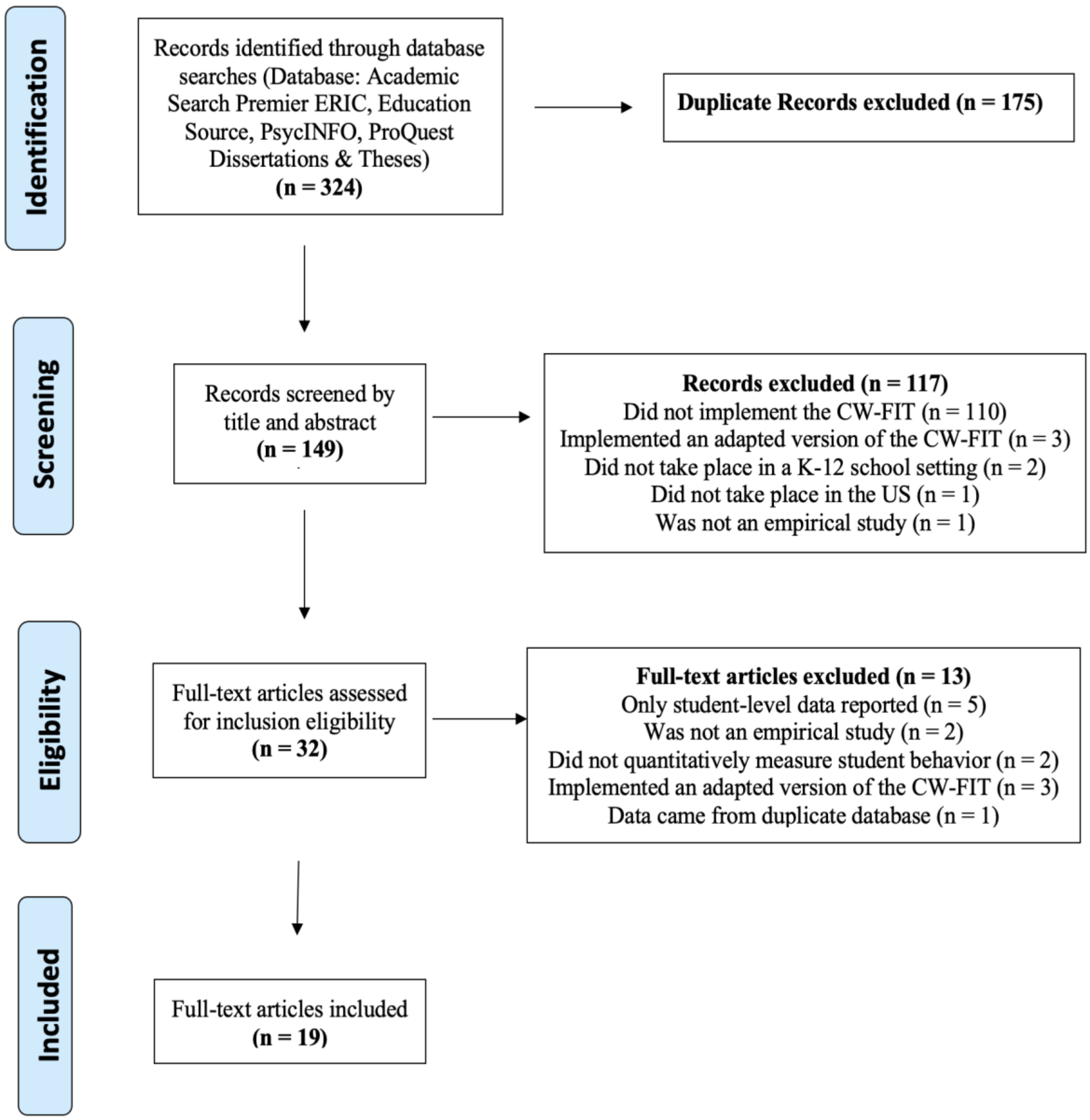
Effectiveness of Tier-1 CW-FIT implementation in classrooms within group design studies

Study	Group	*M* (SD)		Hege's *g*
		Baseline	Intervention	Estimate
Caldarella et al. (2015)	Treatment	*59.79* (13.0)	*74.58* (8.3)	0.52
	Control	*61.63* (13.0)	*69.61* (11.7)	
Kamps et al. (2015b)	Treatment	*51.95* (10.2[†])	*82.99* (8.7[†])	2.94
	Control	*50.18* (10.2[†])	*56.31* (9.5[†])	
Wills et al. (2018)	Treatment	*55.18*[*] (9.6[†])	*79.78*[*] (8.9[†])	2.30
	Control	*57.17*[*] (9.7[†])	*59.17*[*] (9.0[†])	

Note. [*]Grand mean; [†]Converted from SE

Figure 1.1

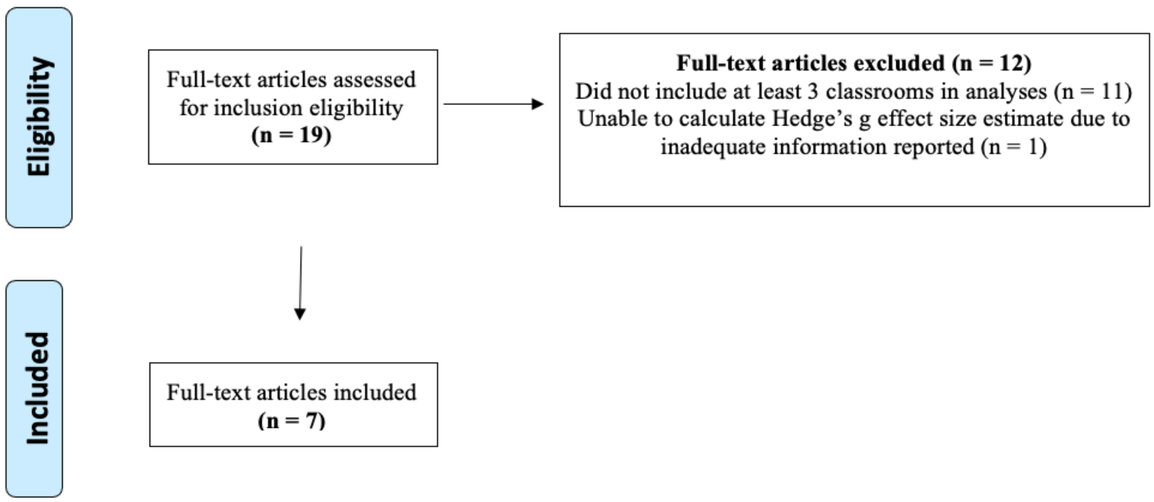
Systematic Review Flow Diagram



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Figure 1.2

Quantitative Synthesis Flow Diagram



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Figure 1.3

Forest Plot

Author	d	se	95% CI: UB	95% CI: LB
Conklin (2017)	2.56	1.26	0.09	5.02
Hansen (2017)	1.51	0.10	1.31	1.70
Kamps (2011)	1.95	0.21	1.53	2.36
Nelson (2018)	2.58	0.37	1.85	3.30
Parikh (2019)	1.35	0.22	0.91	1.78
Speight (2020)	3.77	0.29	3.20	4.33
Wills (2014)	4.65	1.18	2.33	6.96

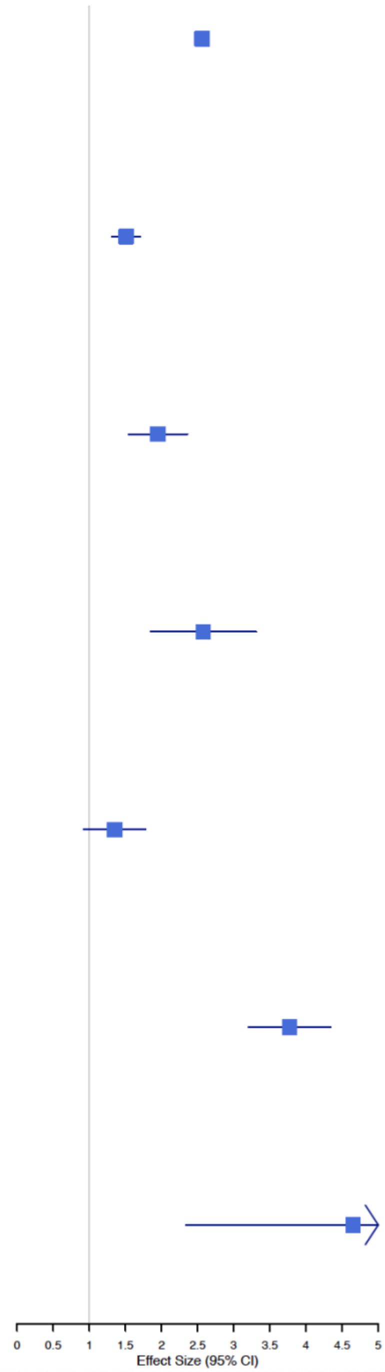


Figure 1.4

Funnel Plot

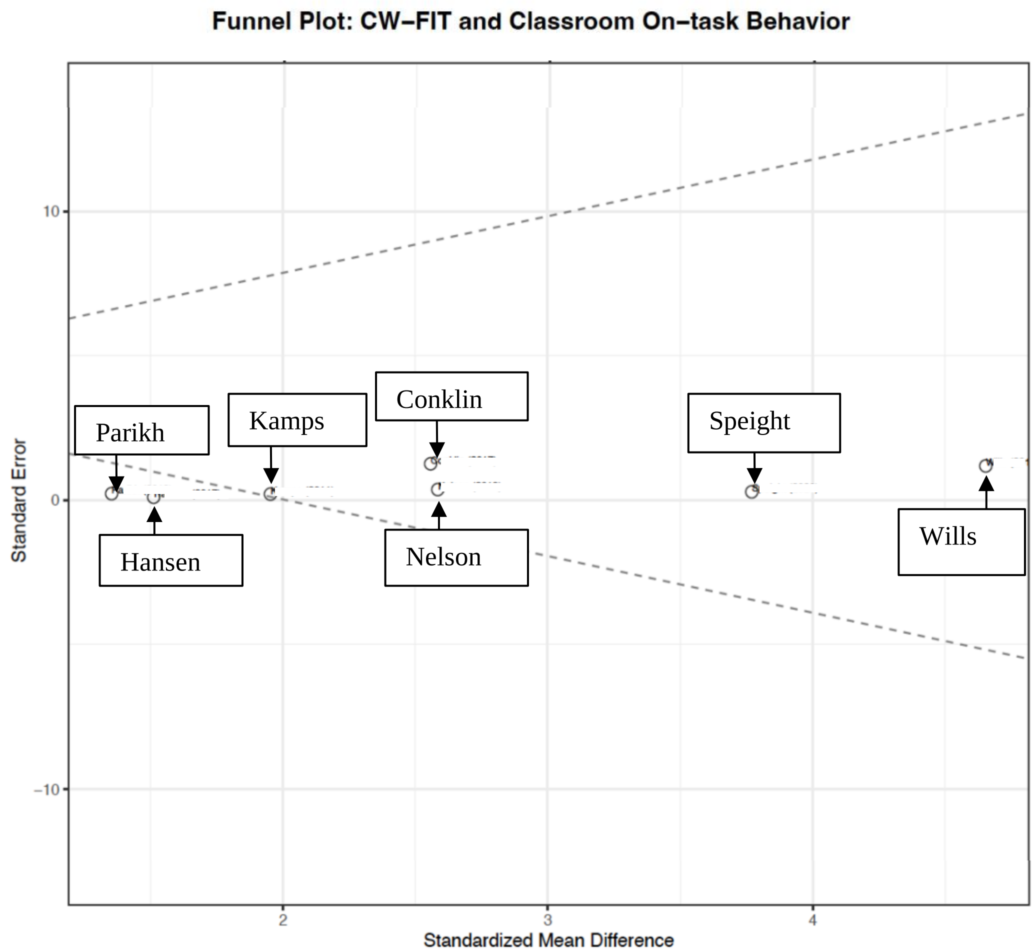


Table 2.1

Enrollment and ODR Characteristics by Racial and Ethnic Group

Racial and Ethnic Group	Minimum % Enrollment	Maximum % Enrollment	Average % Enrollment	Minimum Avg. ODRs	Maximum Avg. ODRs	Avg. ODRs
Asian	0%	18.91%	4.33%	0	6.50	0.35
Black	0%	64.94%	10.26%	0	51.83	1.91
Hispanic	0%	55.56%	22.53%	0	22.00	0.89
Indigenous/Native	0%	74.15%	0.19%	0	12.00	1.66
Multiracial	0%	93.77%	5.53%	0	34.00	1.11
Pacific Islander	0%	29.59%	0.53%	0	11.00	0.35
White	0%	81.40%	54.69%	0	17.17	1.04

Table 2.2

Multiple Linear Regression Models of Modified Total BoQ-R Score for Log-transformed Average ODRs Across Racial and Ethnic Groups

	Avg. ODRs for Asian students	Avg. ODRs for Black students	Avg. ODRs for Hispanic/Latine students	Avg. ODRs for Indigenous/Native students
Intercept	0.41310* (0.11456)	0.47057* (0.24017)	0.42345* (0.19137)	0.19719 (0.26218)
Covariates				
Enrollment by Race and Ethnicity	0.00061 (0.00042)	-0.00041 (0.00036)	-0.00044* (0.00018)	0.00102 (0.00057)
FRL	0.04083 (0.07557)	0.07128 (0.17382)	0.09289 (0.13550)	0.06066 (0.28541)
Middle School	0.04134 (0.04828)	0.19264 (0.10525)	0.15780 (0.08026)	0.29885* (0.14088)
High School	-0.07860 (0.06271)	0.02417 (0.14658)	0.31335* (0.14254)	0.26022 (0.17477)
Total BoQ-R Score (modified)	-0.00283* (0.00121)	0.00501 (0.00260)	0.00167 (0.00205)	0.00143 (0.00297)
Model Summary				
Adjusted R^2	.032	.016	.045	.064

Note. $*p < .05$

Table 2.2 (continued)

Multiple Linear Regression Models of Modified Total BoQ-R Score for Log-transformed Average ODRs Across Racial and Ethnic Groups

	Avg. ODRs for Multiracial students	Avg. ODRs for Pacific Islander students	Avg. ODRs for White students
Intercept	0.17432 (0.18831)	0.41641* (0.21184)	0.49289* (0.23482)
Covariates			
Enrollment by Race and Ethnicity	0.00141 (0.00088)	0.00238 (0.00236)	-
FRL	0.21697 (0.13766)	-0.18885 (0.27553)	-0.14938 (0.27648)
Middle School	0.22900* (0.08655)	0.07303 (0.13161)	0.00376 (0.12402)
High School	0.04480 (0.10845)	0.30826 (0.15631)	0.24604 (0.14829)
Total BoQ-R Score (modified)	0.00155 (0.00203)	-0.00137 (0.00271)	-0.00211 (0.00286)
Model Summary			
Adjusted R^2	.040	.100	.071

Note. *$p < .05$

Table 2.3

Multiple Linear Regression Models of BoQ-R Critical Elements for Log-transformed Average ODRs Across Racial and Ethnic Groups

	Avg. ODRs for Asian students	Avg. ODRs for Black students	Avg. ODRs for Hispanic/Latine students
Intercept	0.51629* (0.14118)	0.46137 (0.28577)	0.36454 (0.22322)
Covariates			
Enrollment by Race and Ethnicity	0.00049 (0.00044)	-0.00041 (0.00036)	-0.00036 (0.00018)
FRL	0.01031 (0.07341)	0.08942 (0.17767)	0.02585 (0.14559)
Middle School	0.03438 (0.04893)	0.18593 (0.10694)	0.11892 (0.08201)
High School	-0.05479 (0.06197)	0.04925 (0.14805)	0.24060* (0.11757)
BoQ-R Critical Systems			
Effective Discipline Procedures (modified)	-0.00812 (0.01919)	0.00349 (0.04035)	0.03226 (0.03043)
Data Entry	0.03452* (0.01616)	0.06657 (0.03403)	0.04049 (0.02661)
Expectations and Rules	-0.02648 (0.01847)	0.01645 (0.04140)	0.00553 (0.03118)
Reward System	0.00518 (0.01170)	-0.00278 (0.02777)	0.00383 (0.01988)
Lesson Plans	-0.00018 (0.01389)	0.04901 (0.02931)	-0.01378 (0.02285)
Implementation Plans	-0.01600 (0.01217)	-0.02598 (0.02824)	-0.00009 (0.02123)
Classroom Systems	0.00527 (0.01100)	-0.02184 (0.02523)	-0.01121 (0.01862)
Evaluation	-0.01774 (0.01646)	0.00174 (0.03251)	-0.01482 (0.02457)
Model Summary			
Adjusted R^2	.054	.025	.034

Note. $*p < .05$

Table 2.3 (continued)

Multiple Linear Regression Models of BoQ-R Critical Elements for Log-transformed Average ODRs Across Racial and Ethnic Groups

	Avg. ODRs for Indigenous/Native students	Avg. ODRs for Multiracial students	Avg. ODRs for Pacific Islander students	Avg. ODRs for White students
Intercept	0.20429 (0.27860)	0.41512 (0.23445)	0.66835* (0.24595)	0.71101* (0.27004)
Covariates				
Enrollment by Race and Ethnicity	0.00089 (0.00061)	0.00152 (0.00091)	0.00201 (0.00229)	-
FRL	-0.10197 (0.17212)	0.16268 (0.14381)	-0.20763 (0.18663)	-0.17006 (0.20018)
Middle School	0.24585* (0.11956)	0.24070* (0.08798)	0.05772 (0.10125)	0.03411 (0.11862)
High School	0.10709 (0.15353)	0.06574 (0.11427)	0.27827* (0.13371)	0.25586 (0.14395)
BoQ-R Critical Systems				
Effective Discipline Procedures (modified)	0.01407 (0.03875)	-0.00682 (0.03219)	0.00344 (0.03647)	0.00176 (0.03680)
Data Entry	0.06757 (0.03664)	0.02672 (0.02968)	0.05832 (0.02993)	0.04874 (0.03754)
Expectations and Rules	0.00780 (0.04004)	-0.05286 (0.03181)	-0.08907* (0.04020)	-0.08660* (0.03979)
Reward System	-0.00878 (0.03049)	0.02513 (0.02176)	0.00577 (0.02833)	0.00461 (0.02904)
Lesson Plans	-0.05526 (0.03463)	-0.00831 (0.02709)	0.02332 (0.03171)	0.03177 (0.03346)
Implementation Plans	-0.02145 (0.02707)	0.00988 (0.02079)	-0.02685 (0.02544)	-0.03167 (0.02926)
Classroom Systems	0.02285 (0.02335)	0.01671 (0.02037)	0.00838 (0.02257)	0.01357 (0.02562)
Evaluation	0.00119 (0.03392)	-0.02224 (0.02810)	0.01054 (0.02692)	0.00772 (0.03091)
Model Summary				
Adjusted R^2	.083	.052	.120	.089

Note. *$p < .05$

Table 2.4

Multiple Linear Regression Models of Modified Total BoQ-R Score for Log-transformed ODR Risk Across Racial and Ethnic Groups

	ODR Risk for Asian students	ODR Risk for Black students	ODR Risk for Hispanic/Latine students
Intercept	0.51677* (0.12253)	1.12963* (0.22397)	1.05544* (0.16600)
Covariates			
Enrollment by Race and Ethnicity	0.00092* (0.00045)	0.00068* (0.00033)	-0.00020 (0.00016)
FRL	-0.03943 (0.07914)	-0.42183* (0.16375)	-0.20195 (0.12486)
Middle School	0.07770 (0.05437)	0.07023 (0.09421)	0.12782 (0.07132)
High School	-0.02374 (0.07383)	-0.09435 (0.12826)	0.20574* (0.09326)
Total BoQ-R Score (modified)	-0.00340* (0.00129)	0.00169 (0.00244)	-0.00305 (0.00178)
Model Summary			
Adjusted R^2	.061	.028	.050

Note. **p* < .05

Table 2.4 (continued)

Multiple Linear Regression Models of Modified Total BoQ-R Score for Log-transformed ODR Risk Across Racial and Ethnic Groups

	ODR Risk for Indigenous/Native students	ODR Risk for Multiracial students	ODR Risk for Pacific Islander students
Intercept	0.73207* (0.24764)	0.66359* (0.17980)	0.66704* (0.20250)
Covariates			
Enrollment by Race and Ethnicity	0.00079 (0.00055)	0.00156 (0.00086)	0.00224 (0.00209)
FRL	-0.21634 (0.20543)	0.08094 (0.14098)	-0.00452 (0.21094)
Middle School	0.20194 (0.13020)	0.11505 (0.08156)	-0.03774 (0.09463)
High School	0.48373* (0.15500)	-0.09317 (0.10616)	0.34203* (0.12461)
Total BoQ-R Score (modified)	-0.00330 (0.00262)	-0.00233 (0.00191)	-0.00495* (0.00215)
Model Summary			
Adjusted R^2	.101	.025	.118

Note. **p* < .05

Table 2.5

Multiple Linear Regression Models of BoQ-R Critical Elements for Log-transformed ODR Risk Across Racial and Ethnic Groups

	ODR Risk for Asian students	ODR Risk for Black students	ODR Risk for Hispanic/Latine students
Intercept	0.54127* (0.14744)	1.02618* (0.26710)	1.04105* (0.20623)
Covariates			
Enrollment by Race and Ethnicity	0.00081 (0.00047)	0.00069* (0.00034)	-0.00019 (0.00017)
FRL	-0.03863 (0.08264)	-0.34271* (0.15537)	-0.28385* (0.13532)
Middle School	0.07829 (0.05166)	0.08968 (0.09803)	0.11876 (0.07653)
High School	-0.00869 (0.07108)	-0.06363 (0.12804)	0.19473 (0.10436)
BoQ-R Critical Systems			
Effective Discipline Procedures (modified)	-0.02519 (0.01974)	-0.02350 (0.03929)	0.00703 (0.02914)
Data Entry	0.00863 (0.01733)	-0.01710 (0.03184)	-0.00695 (0.02492)
Expectations and Rules	0.00641 (0.01922)	0.04891 (0.03816)	0.00573 (0.02925)
Reward System	0.00095 (0.01271)	-0.00049 (0.02414)	0.00595 (0.01831)
Lesson Plans	-0.00497 (0.01461)	0.02728 (0.03280)	-0.02758 (0.02144)
Implementation Plans	0.00559 (0.01356)	-0.00038 (0.02388)	0.01805 (0.01884)
Classroom Systems	-0.00621 (0.01187)	-0.00250 (0.02362)	-0.00249 (0.01691)
Evaluation	-0.01837 (0.01678)	-0.01294 (0.03080)	-0.02940 (0.02186)
Model Summary			
Adjusted R^2	.061	.024	.058

Note. **p* < .05

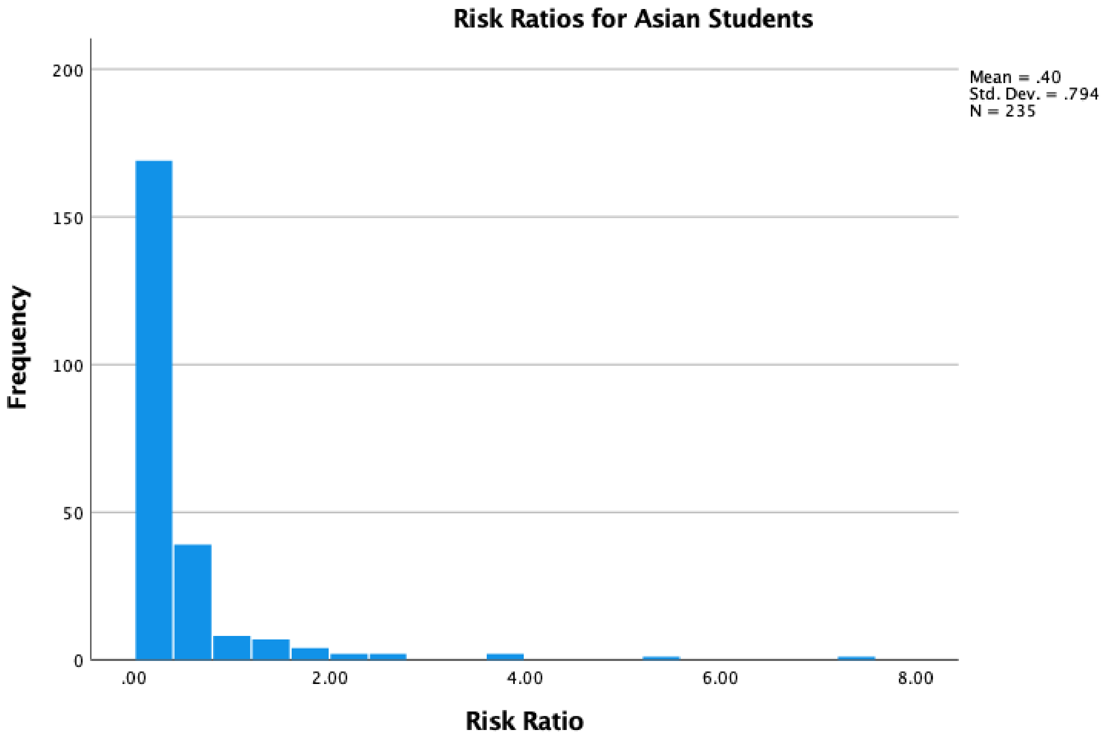
Table 2.5 (continued)

Multiple Linear Regression Models of BoQ-R Critical Elements for Log-transformed ODR Risk Across Racial and Ethnic Groups

	ODR Risk for Indigenous/Native students	ODR Risk for Multiracial students	ODR Risk for Pacific Islander students
Intercept	0.48003 (0.27033)	1.05203* (0.21836)	0.77880* (0.21397)
Covariates			
Enrollment by Race and Ethnicity	0.00068 (0.00058)	0.00101 (0.00082)	0.00244 (0.00203)
FRL	-0.27902 (0.16419)	0.11050 (0.13210)	-0.08550 (0.17547)
Middle School	0.14219 (0.10908)	0.13486 (0.07776)	-0.02019 (0.08920)
High School	0.41019* (0.16400)	-0.06304 (0.10176)	0.39279* (0.12595)
BoQ-R Critical Systems			
Effective Discipline Procedures (modified)	-0.02473 (0.03862)	-0.06287* (0.02982)	0.07359* (0.03433)
Data Entry	0.02025 (0.03793)	-0.03259 (0.02530)	0.03147 (0.02849)
Expectations and Rules	0.04079 (0.03796)	-0.04213 (0.02968)	-0.10252* (0.03392)
Reward System	0.00016 (0.02907)	0.00217 (0.01907)	0.02684 (0.02585)
Lesson Plans	-0.07885* (0.03241)	0.02715 (0.02186)	0.00656 (0.02774)
Implementation Plans	-0.01165 (0.02627)	0.03103 (0.02024)	-0.04412 (0.02271)
Classroom Systems	0.01396 (0.02441)	0.00528 (0.01933)	-0.03097 (0.02098)
Evaluation	0.01447 (0.03361)	-0.00331 (0.02456)	0.01428 (0.02498)
Model Summary			
Adjusted R^2	.133	.060	.223

Note. *$p < .05$

Figure 2.1



Risk Ratios for Asian Students
Mean = .40
Std. Dev. = .794
N = 235
Frequency
200
150
100
50
0
.00
2.00
4.00
6.00
8.00
Risk Ratio

Figure 2.2

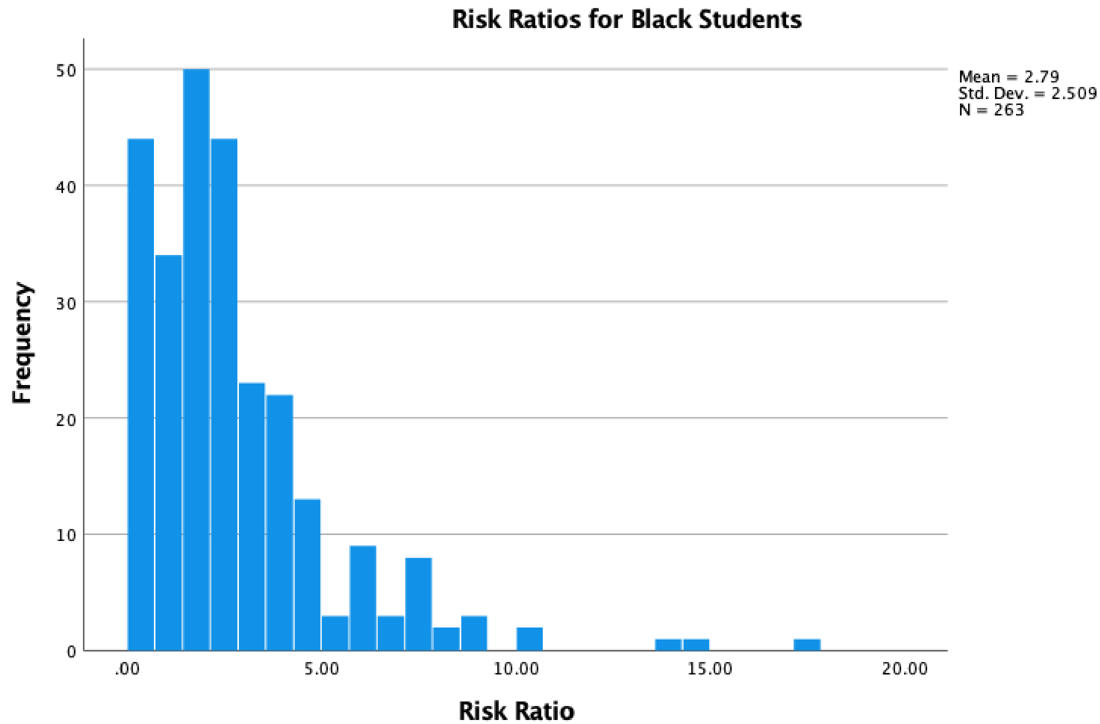

Risk Ratios for Black Students
Mean = 2.79
Std. Dev. = 2.509
N = 263
Frequency
0
10
20
30
40
50
.00
5.00
10.00
15.00
20.00
Risk Ratio

Figure 2.3

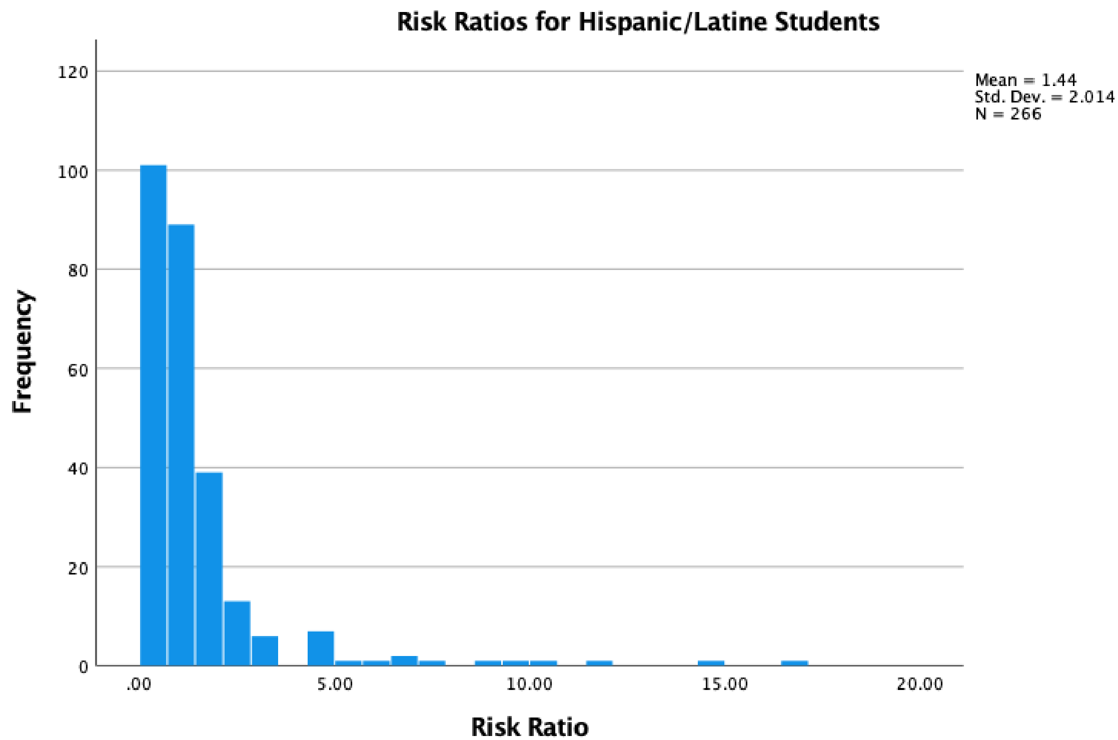

Risk Ratios for Hispanic/Latine Students
Mean = 1.44
Std. Dev. = 2.014
N = 266
Frequency
120
100
80
60
40
20
0
.00
5.00
10.00
15.00
20.00
Risk Ratio

Figure 2.4

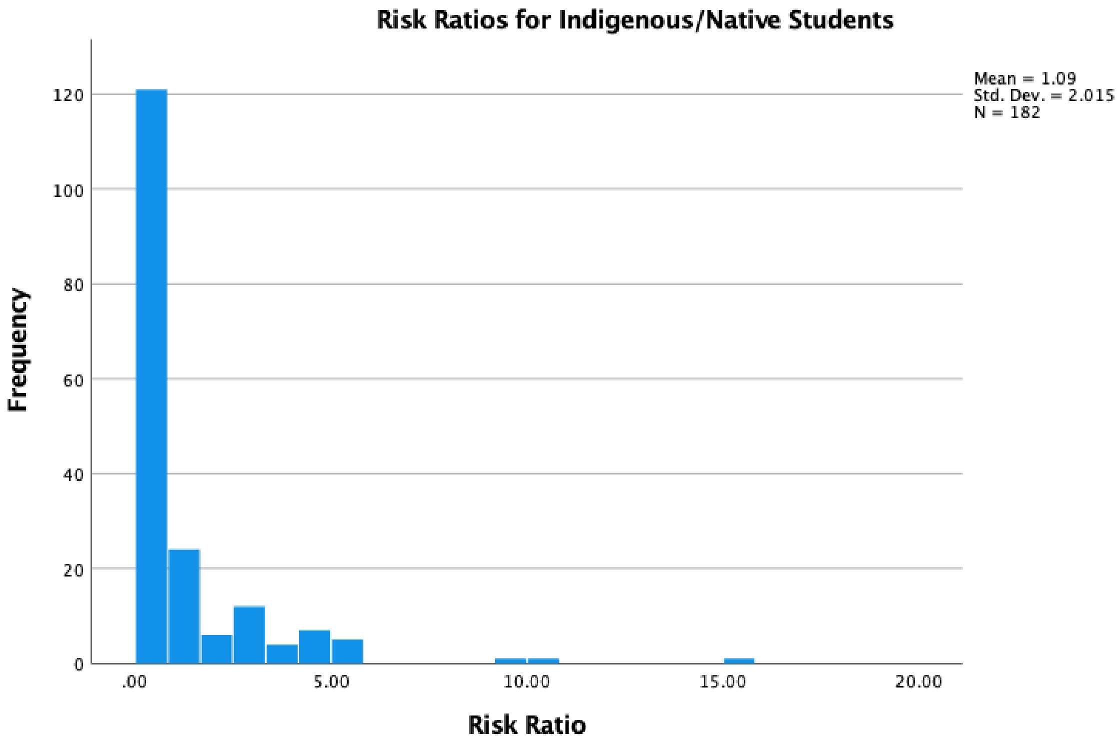

Risk Ratios for Indigenous/Native Students
Mean = 1.09
Std. Dev. = 2.015
N = 182
Frequency
120
100
80
60
40
20
0
.00
5.00
10.00
15.00
20.00
Risk Ratio

Figure 2.5

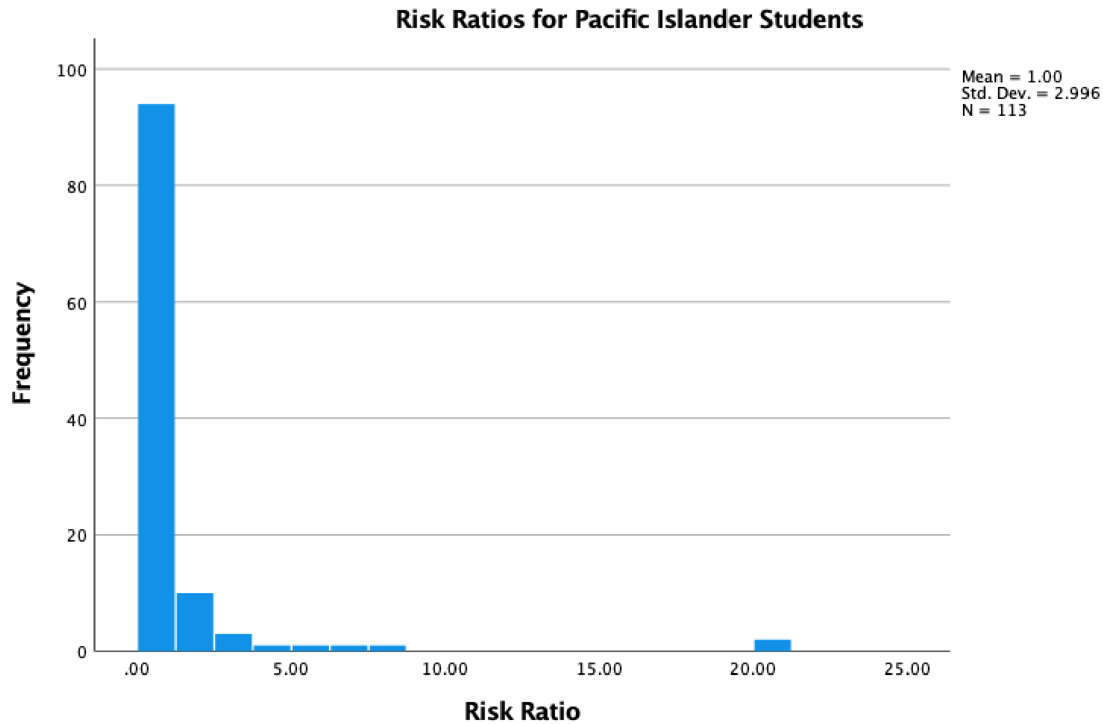

Risk Ratios for Pacific Islander Students
Mean = 1.00
Std. Dev. = 2.996
N = 113
Frequency
100
80
60
40
20
0
.00
5.00
10.00
15.00
20.00
25.00
Risk Ratio

Figure 2.6

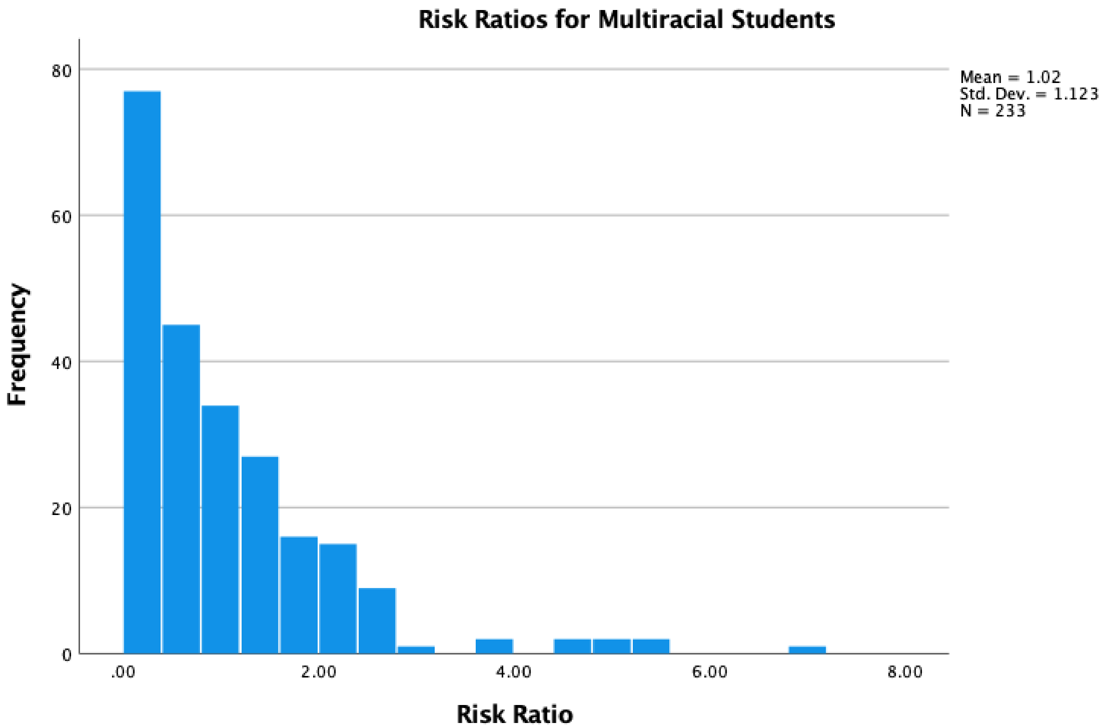
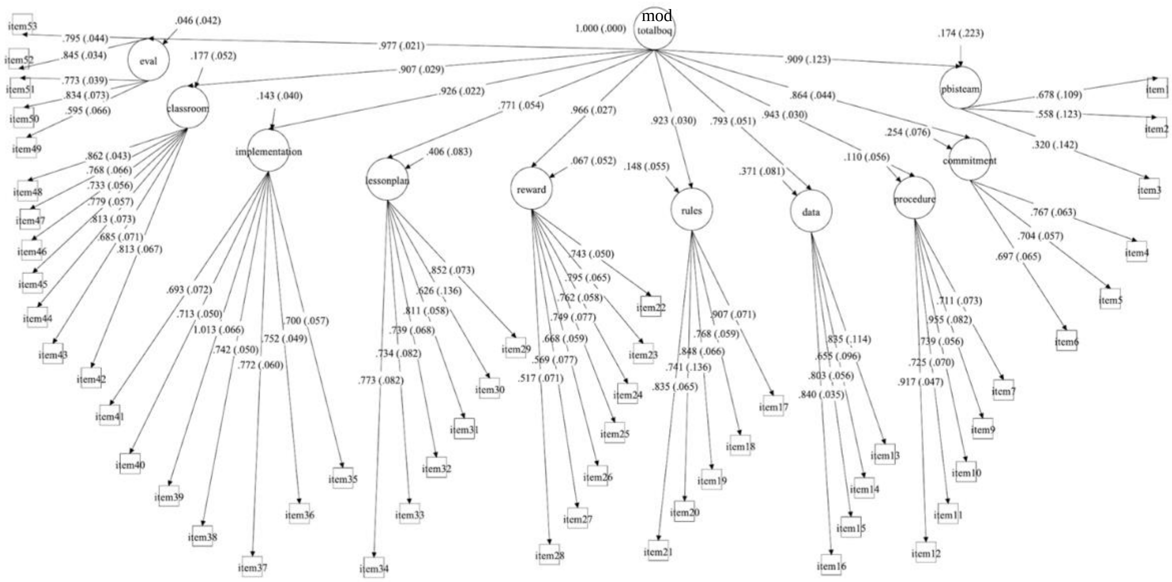

Risk Ratios for Multiracial Students
Mean = 1.02
Std. Dev. = 1.123
N = 233
Frequency
0
20
40
60
80
.00
2.00
4.00
6.00
8.00
Risk Ratio

Figure 2.7

Measurement Model for Modified BoQ-R



References

Algozzine, K., & Algozzine, B. (2007). Classroom instructional ecology and school-wide positive behavior support. *Journal of Applied School Psychology*, *24*(1), 29-47. https://doi.org/10.1300/j370v24n01_02

Anyon, Y., Zhang, D., & Hazel, C. (2016). Race, exclusionary discipline, and connectedness to adults in secondary schools. *American Journal of Community Psychology*, *57*(3-4), 342-352. https://doi.org/10.1002/ajcp.12061

Baker, P. H. (2005). Managing student behavior: How ready are teachers to meet the challenge?. *American Secondary Educatio*n, 51-64.

Barbetta, P. M., Norona, K. L., & Bicard, D. F. (2005). Classroom behavior management: A dozen common mistakes and what to do instead. *Preventing School Failure: Alternative Education for Children and Youth*, *49*(3), 11-19. https://doi.org/10.3200/psfl.49.3.11-19

Barclay, C. M. (2017). *Benchmarks of equality? School-wide positive behavior interventions and supports and school discipline risk and disparities for Black and Hispanic students.* (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.

Barclay, C. M., Castillo, J., & Kincaid, D. (2022). Benchmarks of equality? School-wide positive behavioral interventions and supports and the discipline gap. *Journal of Positive Behavior Interventions*, *24*(1), 4-16. https://doi.org/10.1177/10983007211040097

Bastable, E., Falcon, S. F., Nese, R., Meng, P., & McIntosh, K. (2021). Enhancing school-wide positive behavioral interventions and supports tier 1 core practices to improve disciplinary equity. *Preventing School Failure: Alternative Education for Children and Youth*, *65*(4), 283-290. https://doi.org/10.1080/1045988x.2021.1937020

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics*, *22*(1), 23-36.

Borenstein, M., & Higgins, J. (2013). Meta-analysis and Subgroups. *Prevention Science*, *14*(2), 134-143.

Bottiani, J. H., Bradshaw, C. P., & Gregory, A. (2018). Nudging the gap: Introduction to the special issue "closing in on discipline disproportionality." *School Psychology Review*, *47*(2), 109–117. https://doi.org/10.17105/SPR-2018-0023.V47-2

Bottiani, J. H., Bradshaw, C. P., & Mendelson, T. (2017). A multilevel examination of racial disparities in high school discipline: Black and white adolescents' perceived equity, school belonging, and adjustment problems. *Journal of Educational Psychology*, *109*(4), 532-545. https://doi.org/10.1037/edu0000155

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*(2), 100-115. https://doi.org/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*, 133–149. https://10.1177/1098300709334798

Bradshaw, C. P., Pas, E. T., Bottiani, J. H., Debnam, K. J., Reinke, W. M., Herman, K. C., & Rosenberg, M. S. (2018). Promoting cultural responsivity and student engagement through double check coaching of classroom teachers: An efficacy study. *School Psychology Review*, *47*(2), 118–134. https://doi.org/10.17105/SPR-2017-0119.V47-2

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2012). Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics*, *130*(5), e1136-e1145. https://doi.org/10.1542/peds.2012-0243

Brown-Browner, M. U. L. (2019). *Teachers' Viewpoints on Various Ways They Reduce Discipline Referrals* (Doctoral dissertation, Grand Canyon University).

Bryan, N. (2017). White teachers' role in sustaining the school-to-prison pipeline: Recommendations for teacher education. *The Urban Review*, *49*(2), 326-345. https://doi.org/10.1007/s11256-017-0403-3

*Caldarella, P., Williams, L., Hansen, B. D., & Wills, H. (2015). Managing student behavior with class-wide function-related intervention teams: An observational study in early elementary classrooms. *Early Childhood Education Journal*, *43*(5), 357-365. https://doi.org/10.1007/s10643-014-0664-3

*Caldarella, P., Williams, L., Jolstead, K. A., & Wills, H. P. (2017). Managing student behavior in an elementary school music classroom: A study of class-wide function-related intervention teams. *Update: Applications of Research in Music Education*, *35*(3), 23-30. https://doi.org/10.1177/8755123315626229

Castillo, J. M., Scheel, N. L., Wolgemuth, J. R., Latimer, J. D., & Green, S. M. (2022). A scoping review of the literature on professional learning for MTSS. *Journal of School Psychology*, *92*, 166-187. https://doi.org/10.1016/j.jsp.2022.03.010

Chafouleas, S. M., Sanetti, L. M. H., Jaffery, R., & Fallon, L. M. (2012). An evaluation of a classwide intervention package involving self-management and a group contingency on classroom behavior of middle school students. *Journal of Behavioral Education*, *21*(1), 34-57. https://doi.org/10.1007/s10864-011-9135-8

Childs, K. E., Kincaid, D., & George, H. P. (2011). The revised school-wide PBIS Benchmarks of Quality (BoQ). Available from http://www.pbis.org

Childs, K. E., Kincaid, D., George, H. P., & 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*(2), 89-99. https://doi.org/10.1177/1098300715590398

Children's Defense Fund. (1975). Out-of-school suspensions: Are they helping children? Cambridge, MA: Children's Defense Fund.

Chu, E. M., & Ready, D. D. (2018). Exclusion and urban public high schools: Short-and long-term consequences of school suspensions. *American Journal of Education*, *124*(4), 479-509. https://doi.org/10.1086/698454

Cohen, R., Kincaid, D., & Childs, K. E. (2007). Measuring school-wide positive behavior support implementation: Development and validation of the benchmarks of quality. *Journal of Positive Behavior Interventions*, *9*(4), 203-213. https://doi.org/10.1177/10983007070090040301

*Conklin, C. G., Kamps, D., & Wills, H. (2017). The effects of class-wide function-related intervention teams (CW-FIT) on students' prosocial classroom behaviors. *Journal of Behavioral Education*, *26*(1), 75-100. https://doi.org/10.1007/s10864-016-9252-5

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education*, *36*(4), 220-234. https://doi.org/10.1177/0741932514557271

Cook, C. R., Duong, M. T., McIntosh, K., Fiat, A. E., Larson, M., Pullmann, M. D., & McGinnis, J. (2018). Addressing discipline disparities for black male students: Linking malleable root causes to feasible and effective practices. *School Psychology Review*, *47*(2), 135–152. https://doi.org/10.17105/SPR-2017-0026.V47-2

Cruz, R. A., & Rodl, J. E. (2018). Crime and punishment: An examination of school context and student characteristics that predict out-of-school suspension. *Children and Youth Services Review*, *95*, 226-234. https://doi.org/10.1016/j.childyouth.2018.11.007

De Pry, R. L., & Sugai, G. (2002). The effect of active supervision and pre-correction on minor behavioral incidents in a sixth grade general education classroom. *Journal of Behavioral Education*, *11*(4), 255-267.

Dougherty, E. H., & Dougherty, A. (1977). The daily report card: A simplified and flexible package for classroom behavior management. *Psychology in the Schools*, *14*(2), 191-195. https://doi.org/10.1002/1520-6807(197704)14:2<191::aid-pits2310140213>3.0.co;2-j

Dupper, D. R. (2010). Does the punishment fit the crime? The impact of zero tolerance discipline on at-risk youths. *Social Work in Education*, *32*(2), 67-69. https://doi.org/10.1093/cs/32.2.67

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. Journal of *Pediatric Psychology*, *34*(9), 917-928. https://doi.org/10.1093/jpepsy/jsp004

Effect Size Calculator for T-Test [Social Science Statistics]. Retrieved May 18, 2023, from https://www.socscistatistics.com/effectsize/default3.aspx

Eisenman, G., Edwards, S., & Cushman, C. A. (2015). Bringing reality to classroom management in teacher education. *Professional Educator*, *39*(1), 1-12

Ekholm, E., & Chow, J. (2018). Addressing publication bias in educational psychology. *Translational Issues in Psychological Science, 4*(4), 425. https://doi.org/10.1037/tps0000181

Estrapala, S., Rila, A., & Bruhn, A. L. (2021). A systematic review of Tier 1 PBIS implementation in high schools. *Journal of Positive Behavior Interventions*, *23*(4), 288-302. https://doi.org/10.1177/1098300720929684

Fallon, L. M., McCarthy, S. R., & Sanetti, L. M. H. (2014). School-wide positive behavior support (SWPBS) in the classroom: Assessing perceived challenges to consistent implementation in Connecticut schools. *Education and Treatment of Children*, *37*(1), 1-24. https://doi.org/10.1353/etc.2014.0001

Fenning, P., & Rose, J. (2007). Overrepresentation of African American students in exclusionary discipline the role of school policy. *Urban Education*, *42*(6), 536-559. https://doi.org/10.1177/0042085907305039

Fisher, B. W., & Hennessy, E. A. (2016). School resource officers and exclusionary discipline in US high schools: A systematic review and meta-analysis. *Adolescent Research Review*, *1*(3), 217-233. https://doi.org/10.1007/s40894-015-0006-8

Flannery, K. B., Fenning, P., Kato, M. M., & McIntosh, K. (2014). Effects of school-wide positive behavioral interventions and supports and fidelity of implementation on problem behavior in high schools. *School Psychology Quarterly*, *29*(2), 111-124. https://doi.org/10.1037/spq0000039

Florida PBIS Project: Foundations for Implementation (n.d.). Florida Positive Behavioral Interventions & Support Project A Multi-Tiered System of Supports. Retrieved May 18, 2023, from https://flpbis.cbcs.usf.edu/foundations/PBIS.html

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education*, *37*(2), 106-120. https://doi.org/10.1177/0888406413507002

Gage, N. A., Grasley-Boy, N., Peshak George, H., 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*(1), 50-61. https://doi.org/10.1177/1098300718768208

Gage, N. A., Scott, T., Hirn, R., & MacSuga-Gage, A. S. (2018). The relationship between teachers' implementation of classroom management practices and student behavior in elementary school. *Behavioral Disorders*, *43*(2), 302-315. https://doi.org/10.1177/0198742917714809

Gion, C., McIntosh, K., & Falcon, S. (2022). Effects of a multifaceted classroom intervention on racial disproportionality. *School Psychology Review*, *51*(1), 67-83. https://doi.org/10.1080/2372966x.2020.1788906

Girvan, E. J., Gion, C., McIntosh, K., & Smolkowski, K. (2017). The relative contribution of subjective office referrals to racial disproportionality in school discipline. *School Psychology Quarterly*, *32*(3), 392-404. https://doi.org/10.1037/spq0000178

Gopalan, M., & Nelson, A. A. (2019). Understanding the racial discipline gap in schools. *AERA Open*, *5*(2), 1–26. https://doi.org/10.1177/2332858419844613

Grapin, S. L., & Fallon, L. M. (2021). Conceptualizing and dismantling white privilege in school psychology research: An ecological model. *School Psychology Review*, 1-14. https://doi.org/10.1080/2372966x.2021.1963998

Grasley-Boy, N. M., Reichow, B., van Dijk, W., & Gage, N. (2021). A systematic review of tier 1 PBIS implementation in alternative education settings. *Behavioral Disorders*, *46*(4), 199-213. https://doi.org/10.1177/0198742920915648

Gregory, A., Clawson, K., Davis, A., & Gerewitz, J. (2016). The promise of restorative practices to transform teacher-student relationships and achieve equity in school discipline. *Journal of Educational and Psychological Consultation*, *26*(4), 325–353. https://doi.org/10.1080/10474412.2014.929950

Gregory, A., Cornell, D., & Fan, X. (2011). The relationship of school structure and support to

suspension rates for Black and White high school students. *American Educational Research Journal*, *48*(4), 904-934. https://doi.org/10.3102/0002831211398531

Gregory, A., & Roberts, G. (2017). Teacher beliefs and the overrepresentation of Black students in classroom discipline. *Theory Into Practice*, *56*(3), 187-194.

*Hansen, B. D., Caldarella, P., Williams, L., & Wills, H. P. (2017). Managing student behavior in dual immersion classrooms: A study of class-wide function-related intervention teams. *Behavior Modification*, *41*(5), 626-646. https://doi.org/10.1177/0145445517698418

Heidelburg, K., Rutherford, L., & Parks, T. W. (2022). A preliminary analysis assessing SWPBIS implementation fidelity in relation to disciplinary outcomes of black students in urban schools. *The Urban Review*, *54*(1), 138-154. https://doi.org/10.1007/s11256-021-00609-y

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558. https://doi.org/10.1002/sim.1186

Hilberth, M., & Slate, J. R. (2014). Middle school Black and White student assignment to disciplinary consequences: A clear lack of equity. *Education and Urban Society*, *46*(3), 312-328. https://doi.org/10.1177/0013124512446218

*Hirsch, S. E., Healy, S., Judge, J. P., & Lloyd, J. W. (2016). Effects of an interdependent group contingency on engagement in physical education. *Journal of Applied Behavior Analysis*, *49*(4), 975-979. https://doi.org/10.1002/jaba.328

Hirschfield, P. (2009). Another way out: The impact of juvenile arrests on high school dropout. *Sociology of Education*, *82*(4), 368-393. https://doi.org/10.1177/003804070908200404

Hooper, D., Coughlan, J., & Mullen, M. (2008, September). Evaluating model fit: a synthesis of the structural equation modelling literature. In *7th European Conference on research methodology for business and management studies* (pp. 195-200).

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*(1), 3-12. https://doi.org/10.1177/10983007040060010201

James, A. G., Noltemeyer, A., Ritchie, R., Palmer, K., & University, M. (2019). Longitudinal disciplinary and achievement outcomes associated with school-wide PBIS implementation level. *Psychology in the Schools*, *56*(9), 1512-1521. https://doi.org/10.1002/pits.22282

*Kamps, D., Conklin, C., & Wills, H. (2015a). Use of self-management with the CW-FIT group contingency program. *Education and Treatment of Children*, *38*(1), 1-32. https://doi.org/10.1353/etc.2015.0003

*Kamps, D., Wills, H., Dawson-Bannister, H., Heitzman-Powell, L., Kottwitz, E., Hansen, B., & Fleming, K. (2015b). Class-wide function-related intervention teams "CW-FIT" efficacy trial outcomes. *Journal of Positive Behavior Interventions*, *17*(3), 134-145. https://doi.org/10.1177/1098300714565244

*Kamps, D., Wills, H. P., Heitzman-Powell, L., Laylin, J., Szoke, C., Petrillo, T., & Culey, A. (2011). Class-wide function-related intervention teams: Effects of group contingency programs in urban classrooms. *Journal of Positive Behavior Interventions*, *13*(3), 154-167. https://doi.org/10.1177/1098300711398935

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 309-311. https://doi.org/10.1080/0969594x.2016.1156645

Kern, L., & Clemens, N. H. (2007). Antecedent strategies to promote appropriate classroom behavior. *Psychology in the Schools*, *44*(1), 65-75. https://doi.org/10.1002/pits.20206

Kim, J., McIntosh, K., Mercer, S. H., & Nese, R. N. (2018). Longitudinal associations between SWPBIS fidelity of implementation and behavior and academic outcomes. *Behavioral Disorders*, *43*(3), 357-369. https://doi.org/10.1177/0198742917747589

Kincaid, D., Childs, K., & George, H. (2005). School-wide benchmarks of quality. Unpublished instrument, University of South Florida.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693-2710. https://doi.org/10.1002/sim.1482

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment*, *19*(1), 4-17. https://doi.org/10.1017/brimp.2017.16

Lee, T., Cornell, D., Gregory, A., & Fan, X. (2011). High suspension schools and dropout rates for black and white students. *Education and Treatment of Children*, 167-192. https://doi.org/10.1353/etc.2011.0014

Lenz, A. S. (2013). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development*, *46*(1), 64-73. https://doi.org/10.1177/0748175612456401

Li, C. (2013). Little's test of missing completely at random. *The Stata Journal*, *13*(4), 795-809.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Long, A. C., Miller, F. G., & Upright, J. J. (2019). Classroom management for ethnic–racial minority students: A meta-analysis of single-case design studies. *School Psychology*, *34*(1), 1-13. https://doi.org/10.1037/spq0000305

Maag, J. W. (2001). Rewarded by punishment: Reflections on the disuse of positive reinforcement in schools. *Exceptional Children*, *67*(2), 173-186. https://doi.org/10.1177/001440290106700203

Maggin, D. M., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education*, *38*(6), 353-370. https://doi.org/10.1177/0741932517716900

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: effect size analysis in quantitative research. *CBE—Life Sciences Education*, *12*(3), 345-351. https://doi.org/10.1187/cbe.13-04-0082

Marcucci, O. (2020). Parental involvement and the Black–White discipline gap: The role of parental social and cultural capital in American schools. *Education and Urban Society*, *52*(1), 143-168. https://doi.org/10.1177/0013124519846283

McIntosh, K., Brigid Flannery, K., Sugai, G., Braun, D. H., & Cochrane, K. L. (2008). Relationships between academics and problem behavior in the transition from middle school to high school. *Journal of Positive Behavior Interventions*, *10*(4), 243-255. https://doi.org/10.1177/1098300708318961

McIntosh, K., Ellwood, K., McCall, L., & Girvan, E. J. (2018). Using discipline data to enhance equity in school discipline. *Intervention in School and Clinic*, *53*(3), 146-152. https://doi.org/10.1177/1053451217702130

McIntosh, K., Girvan, E. J., Horner, R. H., & Smolkowski, K. (2014). Education not incarceration: A conceptual model for reducing racial and ethnic disproportionality in school discipline. *Journal of Applied Research on Children: Informing Policy for Children at Risk*, *5*,1–22.

McIntosh, K., Massar, M. M., Algozzine, R. F., George, H. P., Horner, R. H., Lewis, T. J., & Swain-Bradway, J. (2017). Technical adequacy of the SWPBIS tiered fidelity inventory. *Journal of Positive Behavior Interventions*, *19*(1), 3-13. https://doi.org/10.1177/1098300716637193

Mitchell, M. M., & Bradshaw, C. P. (2013). Examining classroom influences on student perceptions of school climate: The role of classroom management and exclusionary discipline strategies. *Journal of School Psychology*, *51*(5), 599-610. https://doi.org/10.1016/j.jsp.2013.05.005

Mizel, M. L., Miles, J. N., Pedersen, E. R., Tucker, J. S., Ewing, B. A., & D'Amico, E. J. (2016). To educate or to incarcerate: Factors in disproportionality in school discipline. *Children and Youth Services Review*, *70*, 102-111. https://doi.org/10.1016/j.childyouth.2016.09.009

Molloy, L. E., Moore, J. E., Trail, J., Van Epps, J. J., & Hopfer, S. (2013). Understanding real-world implementation quality and “active ingredients” of PBIS. *Prevention Science*, *14*(6), 593-605. https://doi.org/10.1007/s11121-012-0343-9

Monson, K. D., Caldarella, P., Anderson, D. H., & Wills, H. P. (2020). Improving student behavior in middle school art classrooms: Initial investigation of CW-FIT tier 1. *Journal of Positive Behavior Interventions*, *22*(1), 38-50. https://doi.org/10.1177/1098300719864704

Morgan, M. A., & Wright, J. P. (2018). Beyond Black and White: Suspension Disparities for Hispanic, Asian, and White Youth. *Criminal Justice Review*, *43*(4), 377–398. https://doi.org/10.1177/0734016817721293

Morrison, G. M., Anthony, S., Storino, M., & Dillon, C. (2001). An examination of the disciplinary histories and the individual and educational characteristics of students who participate in an in-school suspension program. *Education and Treatment of Children*, 276-293. https://doi.org/10.1002/yd.23320019205

Muldrew, A. C., & Miller, F. G. (2021). Examining the effects of the personal matrix activity with diverse students. *Psychology in the Schools*, *58*(3), 515-533. https://doi.org/10.1002/pits.22461

*Naylor, A. S., Kamps, D., & Wills, H. (2018). The effects of the CW-FIT group contingency on class-wide and individual behavior in an urban first grade classroom. *Education and Treatment of Children*, *41*(1), 1-30. https://doi.org/10.1353/etc.2018.0000

*Nelson, M. A., Caldarella, P., Hansen, B. D., Graham, M. A., Williams, L., & Wills, H. P. (2018). Improving student behavior in art classrooms: An exploratory study of CW-FIT Tier 1. *Journal of Positive Behavior Interventions*, *20*(4), 227-238. https://doi.org/10.1177/1098300718762744

Noell, G. H., Duhon, G. J., Gatti, S. L., & Connell, J. E. (2002). Consultation, follow-up, and implementation of behavior management interventions in general education. *School Psychology Review*, *31*(2), 217-234.

Noltemeyer, A., Palmer, K., James, A. G., & Wiechman, S. (2019). School-wide positive behavioral interventions and supports (SWPBIS): A synthesis of existing research. *International Journal of School & Educational Psychology*, *7*(4), 253-262. https://doi.org/10.1080/21683603.2018.1425169

Office for Civil Rights, U.S. Department of Education (2002). *Elementary and Secondary School Civil Rights Compliance Reports*. Retrieved from https://www.govinfo.gov/content/pkg/ERIC-ED480745/pdf/ERIC-ED480745.pdf

*Orr, R. K., Caldarella, P., Hansen, B. D., & Wills, H. P. (2020). Managing student behavior in a middle school special education classroom using CW-FIT tier 1. *Journal of Behavioral Education*, *29*, 168-187. https://doi.org/10.1007/s10864-019-09325-w

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*(3), 354. https://doi.org/10.1037/1082-989x.3.3.354

Pane, D. M., Rocco, T. S., Miller, L. D., & Salmon, A. K. (2014). How teachers use power in the classroom to avoid or support exclusionary school discipline practices. *Urban Education*, *49*(3), 297-328. https://doi.org/10.1177/0042085913478620

*Parikh, D. (2019). Effect of class-wide function-based intervention teams (CW-FIT) and the good behavior game (GBG) intervention with children in a classroom (Master's thesis).

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303-322. https://doi.org/10.1177/0145445511399147

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes. *The Journal of Behavioral Health Services & Research*, *39*(4), 417-433. https://doi.org/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 fidelity scores in relation to student outcomes. *Remedial and Special Education*, *40*(1), 6-15. https://doi.org/10.1177/0741932518805192

Pearman, F. A., Curran, F. C., Fisher, B., & Gardella, J. (2019). Are achievement gaps related to discipline gaps? Evidence from national data. *Aera Open*, *5*(4), 1-18. https://doi.org/10.1177/2332858419875440

Petras, H., Masyn, K. E., Buckley, J. A., Ialongo, N. S., & Kellam, S. (2011). Who is most at risk for school removal? A multilevel discrete-time survival analysis of individual-and context-level influences. *Journal of Educational Psychology*, *103*(1), 223-237. https://doi.org/10.1037/a0021545

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS)*. Baltimore, MD: Brookes.

Pinkelman, S. E., McIntosh, K., Rasplica, C. K., Berg, T., & Strickland-Cohen, M. K. (2015). Perceived enablers and barriers related to sustainability of school-wide positive behavioral interventions and supports. *Behavioral Disorders, 40*(3), 171-183. https://doi.org/10.17988/0198-7429-40.3.171

Pustejovsky, J., Chen, M., & Hamilton, B. (2020). *Scdhlm: Estimate hierarchical linear models for single-case designs.* R package version 0.5.0. University of Wisconsin-Madison. https://jepusto.github.io/scdhlm/

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. *Handbook of Special Education*, 168-186. https://doi.org/10.4324/9781315517698-15

Pustejovsky, E. J., and Swan, M. D. (2019). SingleCaseES: A calculator for single-case effect sizes. R package version 0.4.1. https://CRAN.R-project.org/package=SingleCaseES

Reinke, W. M., Lewis-Palmer, T., & Merrell, K. (2008). The classroom check-up: A classwide teacher consultation model for increasing praise and decreasing disruptive behavior. *School Psychology Review*, *37*(3), 315-332.

Robison, S., Jaggers, J., Rhodes, J., Blackmon, B. J., & Church, W. (2017). Correlates of educational success: Predictors of school dropout and graduation for urban students in the Deep South. *Children and Youth Services Review*, *73*, 37-46.

https://doi.org/10.1016/j.childyouth.2016.11.031

Rohatgi, A. (2012) WebPlotDigitalizer: HTML5 based online tool to extract numerical data from plot images. Version 3. [WWW document] URL https://apps.automeris.io/wpd/ (accessed on May 2023).

Schwarzer, G. (2022). Meta-Analysis in R. *Systematic Reviews in Health Research: Meta-Analysis in Context*, 510-534.

Scott, T. M., Park, K. L., Swain-Bradway, J., & Landers, E. (2007). Positive behavior support in the classroom: Facilitating behaviorally inclusive learning environments. *International Journal of Behavioral Consultation and Therapy*, *3*(2), 223-235. https://doi.org/10.1037/h0100800

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER 2015-002). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123-147. https://doi.org/10.1016/j.jsp.2013.11.005

Shores, R. E., Gunter, P. L., & Jack, S. L. (1993). Classroom management strategies: Are they setting events for coercion?. Behavioral disorders, 18(2), 92-102. https://doi.org/10.1177/019874299301800207

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, 351-380. https://doi.org/10.1353/etc.0.0007

Skiba, R. J., Chung, C. G., Trachok, M., Baker, T. L., Sheya, A., & Hughes, R. L. (2014). Parsing disciplinary disproportionality: Contributions of infraction, student, and school characteristics to out-of-school suspension and expulsion. *American Educational Research Journal*, *51*(4), 640-670. https://doi.org/10.3102/0002831214541670

Skiba, R. J., Horner, R. H., Chung, C. G., Karega Rausch, M., May, S. L., & Tobin, T. (2011). Race is not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review*, *40*(1), 85-107. https://doi.org/10.1080/02796015.2011.12087730

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review*, *34*, 317-342. https://doi.org/10.1023/a:1021320817372

Skiba, R. J., Peterson, R. L., & Williams, T. (1997). Office referrals and suspension: Disciplinary intervention in middle schools. *Education and treatment of children*, 295-315.

Smith, D., Ortiz, N. A., Blake, J. J., Marchbanks, M., Unni, A., & Peguero, A. A. (2021). Tipping point: Effect of the number of in-school suspensions on academic failure. *Contemporary School Psychology*, *25*, 466-475. https://doi.org/10.1007/s40688-020-00289-7

Smolkowski, K., Girvan, E. J., McIntosh, K., Nese, R. N., & Horner, R. H. (2016). Vulnerable decision points for disproportionate office discipline referrals: Comparisons of discipline for African American and White elementary school students. *Behavioral Disorders*, *41*(4), 178-195. https://doi.org/10.17988/bedi-41-04-178-195.1

Solomon, B. G., Tobin, K. G., & Schutte, G. M. (2015). Examining the reliability and validity of the effective behavior support self-assessment survey. *Education and Treatment of Children*, *38*(2), 175-191. https://doi.org/10.1353/etc.2015.0007

Sondel, B., Kretchmar, K., & Hadley Dunn, A. (2022). “Who do these people want teaching their children?” White saviorism, colorblind racism, and anti-blackness in “no excuses” charter schools. *Urban Education*, *57*(9), 1621-1650. https://doi.org/10.1177/0042085919842618

*Speight, R., Kucharczyk, S., & Whitby, P. (2021). Effects of a behavior management strategy, CW-FIT, on high school student and teacher behavior. *Journal of Behavioral Education*, 1-20. https://doi.org/10.1007/s10864-020-09428-9

*Speight, R., Whitby, P., & Kucharczyk, S. (2020). Impact of CW-FIT on student and teacher behavior in a middle school. *Journal of Positive Behavior Interventions*, *22*(4), 195-206. https://doi.org/10.1177/1098300720910133

Sugai, G., O'Keeffe, B. V., & Fallon, L. M. (2012). A contextual consideration of culture and school-wide positive behavior support. *Journal of Positive Behavior Interventions*, *14*(4), 197-208. https://doi.org/10.1177/1098300711426334

Sullivan, A. L., Klingbeil, D. A., & Van Norman, E. R. (2013). Beyond behavior: Multilevel analysis of the influence of sociodemographics and school characteristics on students' risk of suspension. *School Psychology Review*, *42*(1), 99–114. https://doi.org/10.1177/1063426610377329

Sullivan, A. L., Miller, F. G., McKevett, N. M., Muldrew, A., Hansen-Burke, A., & Weeks, M. (2020). Leveraging MTSS to Advance, Not Suppress, COVID-Related Equity Issues. *Communique*, *49*(1), 1-5.

Thier, M., & Beach, P. (2019). Stories we don't tell: Research's limited accounting of rural schools. *School Leadership Review*, *14*(2), 1-14.

Tidwell, A., Flannery, K. B., & Lewis-Palmer, T. (2003). A description of elementary classroom discipline referral patterns. *Preventing School Failure: Alternative Education for Children and Youth*, *48*(1), 18-26. https://doi.org/10.1080/1045988x.2003.10871075

Tillery, A. D., Varjas, K., Meyers, J., & Collins, A. S. (2010). General education teachers' perceptions of behavior management and intervention strategies. *Journal of Positive Behavior Interventions*, *12*, 86–102. https://doi.org/10.1177/1098300708330879

U.S. Bureau of the Census. (1994*). Geographic areas reference manual*. US Department of Commerce, Economics and Statistics Administration, Bureau of the Census.

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development*, *93*(4), 403-411. https://doi.org/10.1002/jcad.12038

Vannest, K.J., Parker, R.I., Gonen, O., & Adiguzel, T. (2016). Single Case Research: web based calculators for SCR analysis. (Version 2.0) [Web-based application]. College Station, TX: Texas A&M University. Retrieved Thursday 18th May 2023. Available from singlecaseresearch.org

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261-293. https://doi.org/10.3102/10769986030003261

Vincent, C. G., Sprague, J. R., & Tobin, T. J. (2012). Exclusionary discipline practices across students' racial/ethnic backgrounds and disability status: Findings from the Pacific Northwest. *Education and Treatment of Children*, *35*(4), 585-601. https://doi.org/10.1353/etc.2012.0025

Vincent, C. G., & Tobin, T. J. (2010). The relationship between implementation of school-wide positive behavior support (SWPBS) and disciplinary exclusion of students from various ethnic backgrounds with and without disabilities. *Journal of Emotional and Behavioral Disorders*, *19*(4), 217–232. https://doi.org/10.1177/1063426610377329

Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of schoolwide positive behavioral interventions and supports on bullying and peer rejection: A randomized controlled effectiveness trial. *Archives of Pediatrics & Adolescent Medicine*, *166*(2), 149-156. https://doi.org/10.1001/archpediatrics.2011.755

Wallace, J. M., Jr., Goodkind, S.,Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among US high school students: 1991–2005. *The Negro Educational Review*, *59*(1–2), 47-62.

Walker, H. M., Hops, H., & Greenwood, C. R. (1981). RECESS: Research and development of a behavior management package for remediating social aggression in the school setting. In *The utilization of* *classroom peers as behavior change agents* (pp. 261-303). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-2180-2_8

Wang, J., & Wang, X. (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

*Weeden, M., Wills, H. P., Kottwitz, E., & Kamps, D. (2016). The effects of a class-wide behavior intervention for students with emotional and behavioral disorders. *Behavioral Disorders*, *42*(1), 285-293. https://doi.org/10.17988/bd-14-12.1

*Wills, H. P., Iwaszuk, W. M., Kamps, D., & Shumate, E. (2014). CW-FIT: Group contingency effects across the day. *Education and Treatment of Children*, *37*(2), 191-210. https://doi.org/10.1353/etc.2014.0016

*Wills, H., Kamps, D., Caldarella, P., Wehby, J., & Romine, R. S. (2018). Class-Wide Function-Related Intervention Teams (CW-FIT): Student and teacher outcomes from a multisite randomized replication trial. *The Elementary School Journal*, *119*(1), 29-51. https://doi.org/10.1086/698818

Wills, H. P., Kamps, D., Hansen, B., Conklin, C., Bellinger, S., Neaderhiser, J., & Nsubuga, B. (2009). The classwide function-based intervention team program. Preventing School Failure: *Alternative Education for Children and Youth*, *54*(3), 164-171. https://doi.org/10.1080/10459880903496230

*Wills, H. P., Wehby, J. H., Caldarella, P., & Williams, L. (2022). Supporting elementary school classroom management: An implementation study of the CW-FIT program. *Preventing School Failure: Alternative Education for Children and Youth*, *66*(3), 195-205. https://doi.org/10.1080/1045988x.2021.2013150

Yang, J., Rahardja, S., & Fränti, P. (2019, December). Outlier detection: how to threshold outlier scores?. In *Proceedings of the international conference on artificial intelligence, information processing and cloud computing* (pp. 1-6).

Young, J. L., & Butler, B. R. (2018). A student saved is NOT a Dollar earned: A meta-analysis of school disparities in discipline practice toward black children. *Taboo: The Journal of Culture and Education*, *17*(4), 95-112.

Appendix A

Description of Tier 1 CW-FIT Procedures*

1. Operationally define target behavior(s)
2. Teach students 3 CW-FIT skills (i.e., how to get the teacher's attention, follow directions, and ignore appropriate behavior)
3. Review CW-FIT skill posters
4. Model examples and nonexamples of CW-FIT skills
5. Provide opportunities for students to practice CW-FIT skills and receive feedback
6. Announce the goal and reward
7. Recognize appropriate student/peer behavior
8. Ignore inappropriate student/peer behavior
9. Implement a group contingency
10. Award a team points when every team member appropriately applies CW-FIT skills
11. Do not award a team points if a team member inappropriately applies CW-FIT skills
12. At the end of the intervention period, count the points and immediately provide reward to team(s) who met the goal

*From Wills et al. (2009)