PREDICTIVE AND CONCURRENT VALIDITY OF THE TIERED FIDELITY INVENTORY (TFI)

by

JERIN KIM

A DISSERTATION

Presented to the Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2019

DISSERTATION APPROVAL PAGE

Student: Jerin Kim

Title: Predictive and Concurrent Validity of the Tiered Fidelity Inventory (TFI)

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Special Education and Clinical Sciences by:

Kent McIntosh	Chair
Rob Horner	Core Member
John R. Seeley	Core Member
Keith Zvoch	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded March, 2019

DISSERTATION ABSTRACT

Jerin Kim

Doctor of Philosophy

Department of Special Education and Clinical Sciences

March 2019

Title: Predictive and Concurrent Validity of the Tiered Fidelity Inventory (TFI)

This study evaluated the predictive and concurrent validity of the Tiered Fidelity Inventory (TFI). Structural equation modeling was applied to test the associations between the TFI and student outcomes. First, a total of 1,691 schools with TFI Tier 1 in 2016-17 and school-wide discipline outcomes in 2015-16 and 2016-17 were targeted, finding a negative association between TFI Tier 1 and differences between African American and non-African American students in major office discipline referrals (ODR) per 100 students per day in elementary schools. A sensitivity test with schools with TFI Tier 1, 2, and 3 was conducted, showing a negative association between TFI Tier 1 and the square root of major ODR rates in elementary schools.

Second, TFI Tier 1 was positively related to proportions of students meeting or exceeding state-wide standards in reading from 1,361 schools with TFI Tier 1 and academic outcomes in 2014-15 and 2015-16. Also, the association between TFI Tier 1 and academic outcomes was found to be stronger when schools implemented SWPBIS for 6 or more years. A sensitivity test with schools with TFI Tier 1, 2, and 3 indicated positive associations between TFI Tier 1 and proportions of students meeting or exceeding state-wide standards in both subjects.

Third, TFI Tier2 was positively associated with the logit of proportions of students with CICO daily points from 570 schools with TFI Tier 2 in 2016-17 and CICO outcomes in 2015-16 and 2016-17. Fourth, correlations between the Evaluation subscale of TFI Tier 1 or 2 and relevant measures in 2016-17 were tested from 2,379 schools. TFI Tier 1 Evaluation was positively correlated with counts of TFI administrations, number of fidelity measures, and counts of viewing SWIS Reports. These correlations were significant except for ODRs by staff. Also, TFI Tier 2 Evaluation was significantly positively correlated with years of SWPBIS implementation, years of CICO-SWIS implementation, and counts of viewing CICO Reports except student period, and negatively with counts of viewing student single period.

These findings were discussed by comparing them with previous research findings, suggesting implications for future research and practice, and addressing research limitations.

CURRICULUM VITAE

NAME OF AUTHOR: Jerin Kim

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Wisconsin, Madison
Ewha Womans University, Seoul

DEGREES AWARDED:

Doctor of Philosophy, Special Education, 2017, University of Oregon
Master of Science, Special Education, 2011, University of Wisconsin
Bachelor of Art, Special Education, 2003, Ewha Womans University

AREAS OF SPECIAL INTEREST:

Positive Behavior Interventions and Supports
Cross-Cultural Implementation of Evidence Based Practice
Research Methods

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, University of Oregon, 2015 – 2017
Special Education Teacher, Seonyoo High School, Seoul, 2007 – Present

GRANTS, AWARDS, AND HONORS:

P. E. O. International Peace Scholarship, 2014 - 2016

PUBLICATIONS:

Kim, J., & Heo, G. H., (2016). The effects of teacher readiness and contextual variables on preschool teachers' perception of problem behaviors. *Korean Journal of Child Care and Education Policy, 10*, 119-148.

McIntosh, K., Kim, J., Mercer, S. H., Strickland-Cohen, M. K., & Horner, R. H. (2015). Variables associated with enhanced sustainability of school-wide positive behavioral interventions and supports. *Assessment for Effective Intervention*, *40*, 184-191.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Statement of Problem

The primary focus of this study is the validation of the school-wide positive behavioral interventions and supports Tiered Fidelity Inventory (TFI; Algozzine et al., 2014), which is a tool to assess the fidelity of implementation (FOI) of schoolwide positive behavioral interventions and supports (SWPBIS). Assessing FOI has been increasingly emphasized in implementation research in recognition that FOI builds a logical bridge between adopting the intervention and expected positive outcomes (Mowbray, Holter, Teague, & Bybee, 2003). In a context of randomized controlled trials, FOI is essential for any comparison of intervention and control groups. Without documenting adherence to fidelity, it is difficult to conclude that the intervention causes even obvious changes in outcomes. Also, in non-research contexts, monitoring and evaluating FOI is strongly encouraged to guide effective implementation of research-validated interventions across various settings (Michie, Fixsen, Grimshaw, & Eccles, 2009). Attention to FOI can inform practitioners or potential intervention providers on the details of intervention to promote their adoption of, preparation for, and ongoing engagement to the implementation of the intervention (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005).

Considering the importance of FOI for ensuring positive intervention outcomes through precise implementation, it is necessary to establish an accurate and efficient evaluation system of FOI (Mowbray et al., 2003). Although there are no common rules on how to assess FOI, researchers (Mowbray et al., 2003; Schoenwald et al., 2011) recommended using diverse methods (e.g., interviews, observations, surveys) and sources (e.g., service providers/recipients, experts, permanent products) to address different aspects of FOI. Each type of FOI measure involves different issues to determine the quality of data obtained from those measures. More importantly, using them for actual decision making in either research or non-research contexts requires scientifically sound evidence for reliability and validity so that the decisions driven by the interpretation of obtained scores can be empirically supported (Cizek, 2012).

FOI measures can be validated via testing content validity (Bloomfield, 2015; Sullivan, 2015), internal consistency reliability (Naoom, 2014; Sullivan, 2015), test-retest or inter-rater reliability (Bradshaw, Debnam, Koth, & Leaf, 2009; Cohen, Kincaid, & Childs, 2007; Sullivan, 2015), factorial structure validity (Bassett, Stein, Rossi, & Martin., 2016; Sullivan, 2015), convergent validity (Naoom, 2014), discriminant validity (Bassett et al., 2016; Stefancic, Tsemberis, Messeri, Drake, & Goering, 2013), concurrent validity (Bloomfield, 2015; Cohen et al., 2007), and predictive validity (Forgatch, Patterson, & DeGarmo, 2006). Although these examples pertain to different FOI measures, multiple validation studies for a particular measure are not separately processed. Each type of validation study addresses one part of the argument that justifies a sequence of inferences connecting the observations to the score interpretation and use (Kane, 2013).

From the current edition of *Standards for Educational and Psychological Testing* published by the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (= *Standards*; 2014), validity refers to “the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (p. 14), aligning to Messick's (1989) conceptual view to consider validity (so called construct validity) as a unified but multi-faceted construct (Cizek, 2012; Thorndike & Thorndike-Christ, 2010). This definition differs from the traditional view in a sense that different types of validity (e.g., content-, narrowly defined construct-, criterion-based validity) together build stronger, but not complete, justification of score interpretation and test use (Thorndike & Thorndike-Christ, 2010). Notably, despite ongoing debates, considerations of not only score interpretation but also test use as well as consequences of both were embraced as a source of validity evidence so that two crucial and inter-related validation endeavors for validity of score inferences and justification of test use can be pursued in parallel (Cizek, 2012).

Likewise, *Standards* defined validation as an ongoing process of "constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (AERA, APA, NCME, 2014, p. 9). This definition implies that validation is a continuous, accumulative, but demanding process based on a wide array of multiple studies conducted across different settings and over time. Repeated investigations of multi-faceted validity generate an ongoing, recursive decision-making process by informing the gap between the proposed score interpretations and use and the validation evidence for them and reexamining the original inferential claims, which is the greatest benefit from validation efforts (Cizek, 2012). Also, there is a common recognition that validation is a value-laden process (Cizek, 2012; DeLuca, 2011). Specifically, choices of particular sources of evidence instead of other sources or weighting various sources of evidence differently are implicitly or explicitly affected by beliefs, assumptions, and preferences. More obviously, consideration of values is visible in and often becomes the object of justification of test use (i.e., feasibility, cost-efficiency, utility; Cizek, 2012). With a growing emphasis on test use justification, Cizek recommended clearly articulating values or preferences with relation to evaluating validity arguments.

Kane (2013) pointed out that Messick's definition of validity offers a general and comprehensive framework for validation. However, it does not prescribe specific procedures. In this respect, many researchers (Briggs, 2004; Cizek, 2012; Kane, 2013) proposed an argument-based approach to validation, which describes a relatively simple procedure to (1) define a network of logical claims regarding the score inference and use and (2) evaluate them. The former process pertains to an interpretative argument, which explicitly articulates the proposed score interpretation and use by specifying a series of inferences and assumptions logically relating observed scores to score-based decisions (Kane, 2006). The latter is based on the validity argument, an evaluation of the interpretative argument in which relevant inferences are warranted by the backing obtained from systematically developed rubrics, pilot trials, or empirical studies.

This approach offers a useful structure to evaluate validity evidence for both score interpretations and test use (Chapelle, Enright, & Jamieson, 2010; Cizek, 2012). To build stronger validity evidence, collections and synthesis of all relevant evidential studies are needed to address the discrepancies between the originally specified and the empirically evaluated interpretative arguments (Cizek, 2012). With the continued validation endeavor, selection of sources of validity evidence and appropriate analytic methods for the upcoming validation may depend on not only the progress of the existing research but also the nature of the proposed score interpretation and use, and seriousness of erroneous inference, as well as resources, data and samples available at hand (Cizek, 2012).

Purpose of the Study

This study aims to enhance the validity evidence on TFI score interpretation and use. The TFI is contemporary, comprehensive, and the only measure to assess FOI of SWPBIS across all three tiers (Algozzine et al., 2014). Researchers (Massar et al., in press; McIntosh et al., 2017) investigated technical adequacy of the TFI during the initial validation process and refined the measure for stronger validity and usability. Specifically, they found strong construct validity, interrater and test-retest reliability, feasibility, and concurrent validity with relation to the existing fidelity measures as well as factorial structural validity. Along with a carefully designed development process, the initial validation studies offered internal validity evidence (suggesting that the TFI measures what they intend to measure in an accurate and valid manner) to a score interpretation on the TFI that higher TFI scores indicate stronger implementation of core features of SWPBIS in schools (McIntosh et al., 2017). However, the theoretical and empirical nature of the construct suggests assumptions relating to specific uses of the score interpretation within school contexts where SWPBIS is introduced. Specifically, school teams regularly monitor FOI of SWPBIS based on their belief that the properly implemented core features of SWPBIS improve student learning and behaviors. Likewise, in multiple research studies (e.g., Flannery, Fenning, Kato, & McIntosh, 2014; Freeman et al., 2016), student outcomes were shown to vary as a function of fidelity scores. In a recognition that FOI of SWPBIS as the construct to be measured by the TFI is theoretically and empirically expected to affect student outcomes, the purpose of this study was to evaluate the predictive validity of TFI scores to assess the association between TFI scores and intended student outcomes. Noting that one critical component of SWPBIS, measured by the TFI, relates to regular monitoring of student outcomes or fidelity data, this study will evaluate the concurrent validity of TFI scores to test the association between TFI subscale scores (e.g., the Evaluation Subscale) and other relevant measures regarding data use activity or other evaluation activities. Considering that the TFI is designed to assess FOI of SWPBIS at each of three tiers of support, this study might provide empirical evidence for the logic of the multi-tiered systems of support approach to behavior through testing the associations between tiered fidelity scores on the TFI and behavioral outcomes from schools implementing all three tiers.

Literature Review

Schoolwide Positive Behavioral Interventions and Supports

SWPBIS is a three-tiered framework to deliver preventive and positive behavior support at a systems level (Sugai & Horner, 2009). Multiple randomized control trials (Bradshaw, Mitchell, & Leaf, 2010; Horner et al., 2009) have shown positive effects of SWPBIS, including a reduction in office discipline referrals (ODRs), an improvement in academic achievement, and an increase in student perceptions of school safety. Research (Bradshaw et al., 2010; Horner et al., 2009; Flannery et al., 2014; Freeman et al., 2016) also documented the association between FOI of SWPBIS and improved student outcomes, which provided empirical evidence of the underlying mechanism that the effects of SWPBIS are related to its FOI. For instance, Flannery et al. (2014) found that SWPBIS decreased ODR rates in high schools over three years, and the degree of reduction was significantly related to stronger FOI.

Prevention and Tiered Support System. SWPBIS emphasizes a whole school prevention approach to build a positive, predictable, and safe environment. School-wide efforts begin with defining socially valued outcomes. To achieve these outcomes, teams select a small number of evidence-based interventions and strategies with strong effectiveness and organize them within a multi-tiered framework that differentiates the scopes and intensity of behavior supports by student needs. Multi-tiered systems of support (MTSS) operate based on the prevention logic that high quality interventions and instructions can be first delivered to all students before increasingly intensive supports are introduced to some or successively fewer students (Anderson & Borgmeier, 2010; Walker et al., 1996). To allow students at risk to readily access appropriate interventions within MTSS, systematic screening procedures (i.e., teacher-request for assistance, early identification via school wide screening, matching interventions to student needs) must be clearly defined and set in place along with referral processes before implementation (Anderson & Borgmeier, 2010; Lane, Oakes, Menzies, Oyer, & Jenkins, 2013).

Implementation of Core Features. Behavioral science offers a theoretical mechanism to define core features of an environment needed to change student behaviors (Horner & Sugai, 2015). To illustrate, core features of Tier 1 involve: (a) defining and teaching school-wide behavioral expectations across different settings (i.e., classrooms, cafeteria, library) within a school; (b) implementing a reinforcement system to promote student adherence to behavioral expectations; (c) establishing a fair and consistent consequence system to discourage, correct, and redirect problem behaviors; and (d) regularly collecting and using data for decision making (Horner & Sugai, 2015). All these features must extend to classroom routines so as to enhance consistency in school-wide discipline systems (Reinke, Herman, & Stormont, 2013).

Effective implementation of Tier 1 supports is expected to produce positive outcomes for at least 80% of the school population, whereas the students whose behaviors do not respond to Tier 1 receive Tier 2 supports. Tier 2 is not a replacement of but rather is additional to Tier 1 (Horner & Sugai, 2015). Anderson and Borgmeier (2010) noted that Tier 2 involves (a) explicit instruction and reviews of expected behaviors and relevant skills (i.e., prosocial skills, self-regulation skills), (b) regular and frequent opportunities to practice skills and to receive feedback, (c) structured antecedent prompts for behavioral expectations, (d) fading supports in response to student progress if appropriate, and (e) often offering a tool for communication with parents. One crucial feature of Tier 2 is to establish multiple ongoing practices sharing critical features (e.g., Check in Check out, social skill groups, tutoring), and each intervention is similarly delivered to multiple students readily upon school referrals (McIntosh, Campbell, Carter, & Dickey, 2009). Such a moderate level of intensity of supports, a mixture of standardization and differentiation, enhances cost-effectiveness so that Tier 2 can support a large proportion of students with limited resources (Anderson & Borgmeier, 2010; McIntosh, Campbell, Carter, & Dickey, 2009).

Implementation of Tier 2 supports is intended to provide positive outcomes to 10 to 15% of students, and then 5% or less of the school population, whose behaviors are non-responsive to both Tier 1 and 2, can benefit from the most intensive and individualized supports at Tier 3 (Horner & Sugai, 2015). Tier 3 supports involves in-depth formal assessments, comprehensive and individualized support plans, and management by personalized teams tailored to student needs (Horner & Sugai, 2015). The foundation of Tier 3 is functional behavior assessment (FBA), a process to collect information about student behavior and relevant environmental events, develop a hypothesis regarding the functions and conditions of the problem behavior, and produce a behavior support plan (Scott & Eber, 2003). However, Scott and Eber noted the complexity of student needs at the highest levels may require wraparound process. Wraparound process is characterized as coordination of extensive caring system involving both organizing natural and formal supports, person-centered planning, individualized team, active family and student involvement, and inter-agency collaboration (Eber, Breen, Rose, Unizycki, & London, 2008; Scott & Eber, 2003). Noteworthy is the shared commitment by a student, family, and teachers on a proactive team process toward improving student quality of life (Eber et al., 2008).

Implementation of Support Systems. The SWPBIS framework also highlights systems approach to support and sustain implementation of intervention practices and outcomes (Horner & Sugai, 2015). First, the implementation process is driven by school teams at each tier. Tier 1 team members represent an entire school across different grade levels, subjects, and services, whereas Tier 2 and 3 teams involve one coordinator who oversees the overall implementation process, student support staff, administrators, teachers, and if necessary, other individuals engaging in interventions (Anderson & Borgmeier, 2010; Eber et al., 2008). Due to intensity and complexity, Tier 3 supports require expertise in behavioral intervention strategies and school contexts (Eber et al., 2008). To promote effective and efficient teaming process, it is necessary for school teams to document specific roles and operational procedures, receive sufficient training opportunities and provide administrative support to secure time and costs needed for team activities (Horner, Sugai, Todd, & Lewis-Palmer, 2005; Horner & Sugai, 2015).

Also, professional development is an essential strategy to promote change in teacher behaviors to enhance implementation and outcomes (McIntosh, Filter, Bennett, Ryan, & Sugai, 2010). School teams of each tier document specific implementation plans for professional development so that the entire school staff, team members, and individuals engaging in intervention practices can receive not only training regarding the rationale of SWPBIS and relevant theories and skills needed to demonstrate any particular intervention practices, but also ongoing performance feedbacks.

To guide decision making, school teams should establish a reliable data system that facilitates data gathering, summary, and visual display from different sources of data (e.g., ODRs, attendance, grades, fidelity; Horner et al., 2005). The collected data can be regularly reviewed and used for active decision making regarding planning and implementation strategies. Most of all, well-established data systems support evaluation and ongoing self-monitoring, which allows the existing systems and practices to adapt to, improve, and sustain in ever-changing circumstances (Horner & Sugai, 2015).

Measuring Fidelity of Implementation of SWPBIS

SWPBIS focuses on core features of an environment theoretically presumed to produce intended outcomes rather than stipulating on specific practices. To confirm the extent of implementation of the core features, FOI is commonly measured in a systematic and regular manner (Horner & Sugai, 2015). Thus, defining FOI for a particular intervention cannot proceed without operationalizing not only the nature of FOI but also the essential components of the intervention.

Definitions of Fidelity. There are many synonymous terms for fidelity (e.g., treatment integrity, procedural reliability, treatment adherence/compliance; Gresham, 2009). According to Bellg et al. (2004), a seminal introduction of treatment fidelity was first found in Moncher and Prinz's (1991) article by adding the concept of "differentiation," referring to whether the treatments differ from one another in the intended manner (Kazdin, 1986; Moncher & Prinz, 1991; Waltz et al., 1993) to the traditional definition of integrity as the degree to which the treatment was delivered as originally intended (Moncher & Prinz, 1991). Sharing the same idea that the delivered intervention should be consistent with the intended interventions (Century, Rudnick, & Freeman, 2010), the concept of fidelity or integrity has evolved across various fields, such as medical and health science, rehabilitation, clinical psychology, or applied behavior analysis, as well as education (Gresham, 2009).

Dimensions of fidelity. Ongoing efforts to conceptualize and characterize fidelity indicate that fidelity is a multi-dimensional and complex construct (Sanetti & Kratochwill, 2009). Although there are multiple, overlapping frameworks, many studies (Century et al., 2010; Sanetti & Kratochwill, 2009) cited one study of literature review conducted by Dane and Schneider (1998). They examined outcome studies of prevention programs at primary and early secondary levels published between 1980 and 1994. They found that only 24% (39) of the 162 studies documented evaluation of program integrity, and out of them, at least one of five aspects of integrity were measured, including adherence, exposure, quality of delivery, responsiveness, and program differentiation.

Firstly, adherence was defined as "the extent to which specified program components were delivered as prescribed in program manuals" (Dane & Schneider, 1998,

p. 45). This definition appears almost identical to the broad definition of fidelity (Century et al., 2010). Century et al. argued that adherence should not be an aspect of fidelity, but an equivalent construct in a sense that adherence can be measured for different aspects of fidelity. However, noteworthy is the fact that adherence alone cannot explain a complete definition of fidelity unless essential features of interventions, intervention providers, and recipients are specified. Literally, adherence can be viewed as the relationship between intervention providers and an originally designed intervention (intervention providers are expected to follow the original design of the intervention), which offers a critical element of definition of fidelity. Moreover, as indicated by Dusenbury et al. (2003), operationalizing fidelity as strict adherence to the specified intervention components that conform to theoretical principles become more useful when it comes to discussing adaptation to suit local needs.

Focusing on service delivery, exposure (or dosage) refers to the quantity of service delivery including number of sessions, lengths, or frequency, whereas quality of delivery is defined as "a measure of qualitative aspects of program delivery that are not directly related to the implementation of prescribed content" (Dane & Schneider, 1998, p. 23), or more specifically as "the extent to which a provider approaches a theoretical ideal in terms of delivering program content" (Dusenbury et al., 2003, p. 244) such as implementer attitudes toward program, or leader preparedness. This aspect of fidelity appears relevant to prospective competency of service providers (Dusenbury et al., 2003; Waltz, Addis, Koerner, & Jacobson, 1993). With an emphasis on qualification of service provider, many researchers (e.g., Resneck et al., 2005; Sanetti & Kratochwill, 2009;

Waltz et al., 1993) proposed competence or training of interventionists as a facet of fidelity.

In addition, responsiveness (or participant responsiveness) is understood as the extent of "participant response to program sessions…such as levels of participation and enthusiasm" (Dane & Schneider, 1998, p. 45) or "the extent to which participants are engaged by and involved in the activities and content of the program" (Dusenbury et al., 2003, p. 244). This aspect of fidelity considers that intervention is a dual process between intervention providers and recipients (Centry et al., 2010). For example, if discussion is a critical component of a particular instructional intervention, both teacher facilitation of student discussion and students' active engagement in discussion can be separately measured (Century et al., 2010). Likewise, Bellg et al. (2004) and Resnick et al. (2005) proposed not only receipt of intervention, defined as "the degree to which participant understands and demonstrates knowledge of and ability to use treatment skills" (Bellg et al., 2004, p. 444), but also enactment of interventions, referring to "the degree to which the participant applies the skills learned in treatment in his or her daily life" (p. 444). For example, in a study on instruction of self-regulation skills for writing performance, researchers need to verify that students fully understand self-regulation skills during training and also apply them in their classrooms so that the intervention effect on the purposed outcome can be evaluated.

Lastly, differentiation is defined as "a manipulation check that is performed to safeguard against the diffusion (unintentional spread) of treatments, that is, to ensure that the subjects in each experimental condition receive only the planned intervention" (Dane & Schneider, 1998, p. 45). Other researchers also defined differentiation as unique features of interventions which distinguish one intervention from another (Bellg et al., 2004; Century et al., 2010; Dusenbury et al., 2003). This aspect of fidelity can be emphasized to enhance the intervention to improve intended outcomes (Dusenbury et al., 2003), but measurement of differentiation requires specification of unique components of the intervention distinguishable from other interventions (with similar purposes) or business as usual conditions. Century et al. (2010) considered differentiation as an analytic process under which critical components, either common or differential, are comparatively identified across interventions.

Those multi-dimensions of fidelity are often divided into two aspects, structure and process (Mowbray et al., 2003). Structure was defined as the framework for service delivery (i.e., staffing levels and characteristics, case load size, budget, frequency of contacts), and process refers to the way in which services are delivered (i.e., teacher-student interactions, individualization of treatment, teacher/school climates). Century et al. (2010) related structure to the surface design of the intervention, and process to the participants' behaviors and interactions during the intervention. Mowbray et al. noted that despite its feasibility in measurement and manipulation, only assessing the structure aspect of fidelity might ignore less invisible but more critical components of intervention.

Although comprehensively assessing all different aspects of fidelity for a given intervention can provide a complete picture of fidelity (Dane & Schneider, 1998; Dusenbury et al., 2003), complexity, or specificity, is another critical consideration for effective and efficient measurements. Schoenwald et al. (2011) recommended to focus on the purposes, or intended uses, of the fidelity measure. For instance, if multiple purposes are expected by different stakeholders for a given measure, relevant aspects of fidelity can be specified with reasonably less specificity but wider comprehensiveness to make it feasible and applicable to all users. Also, determination of which aspects of fidelity can be measured is related to whether they are relevant and also crucial to implementing core components of the intervention built based on the theoretical principles, and different components may need assessment of different aspects of fidelity (Schoenwald et al., 2011).

Adherence to Core Components of the Intervention. Once the purposes of an FOI measure are determined, operationalizing FOI requires defining essential components of a particular intervention (Schoenwald et al., 2011). A component is defined as "the major operational features or parts of any innovation" (Hord, Rutherford, Huling-Austin & Hall, p. 13). However, all specified components of the intervention might not be essential to improve outcomes. There are many factors which are not theoretically essential but (directly or indirectly) influential to implementation and outcomes (called mediators or moderators; Sanetti & Kratochwill, 2009; Schoenwald et al., 2011). For example, student involvement may moderate the effects of SWPBIS on reductions in problem behaviors. Whether any potential component is purely contextual or a part of the intervention can be determined depending on the theoretical boundaries of the intervention depicted by the intervention developers or the agreements among experienced implementers (Century et al., 2010).

Behavioral change interventions are designed based on a theory-driven logic of change under which key constructs causally related to the behavioral change are identified (Nelson et al., 2012). Specifically, the mechanism of change links the intervention techniques to outcomes passing through multiple mediators or under the interactions with moderators (Michie et al., 2009; Nelson et al., 2012). Those specified constructs determine core components of intervention, and each component needs to be operationalized and translated to quantitative indicators. However, critical constructs of the logic model might be different across various fields, which have unique service delivery models tailoring to the target recipients (Sanetti & Kratochwill, 2009). Despite some variations in service delivery models across fields, Fixsen, Blase, Naoom, and Wallace (2009) noted that high fidelity behaviors of intervention providers are commonly driven by core implementation components involving staff selection, preservice/in-service training, consultation/coaching, staff performance assessment, decision support data systems, administrative supports, and systems interventions. Therefore, core components of behavioral change intervention need to define core implementation components as a part of the intervention, which operates independently but interactively with contents of intervention, directly delivered to its recipients (Carroll et al., 2007; Fixsen et al., 2009).

Uses of FOI Assessment in SWPBIS. The key consideration for designing a particular instrument is the purpose of the instrument to answer what decisions will be made using the resulting scores from the FOI measure (Schoenwald et al., 2011). There are multiple measures of FOI of SWPBIS, with different levels of precision matching to the relevant decisions on the basis of FOI scores (Algozzine et al., 2010). Specifically, the SWPBS Evaluation Blueprint (Algozzine et al., 2010) suggested three primary purposes of the existing fidelity measures: research, annual evaluation, and progress monitoring.

First, the primary goal of evaluating FOI in efficacy or effectiveness trials is related to internal validity as documenting high fidelity in adherence to theories of change that confirm the functional relations between intervention and resulted outcomes. Even if the intervention fails to improve outcomes, evaluation of FOI allows researchers to explain and address the possible causes of failures and to guide ongoing revisions. Such repeated trials may help determine essential components of the intervention (Mowbray et al., 2003; O'Donnell, 2008). Relevantly, variations in fidelity across different settings within effectiveness trials can inform the possible flexibility or feasibility in FOI, which enhances external validity of the intervention (O'Donnell, 2008). At the end, evaluating FOI in effectiveness trials can guide decisions regarding scale up: proceeding to scale-up (only if high levels of the FOI and positive outcomes were documented), improving FOI (if low levels of the FOI were documented), or re-designing the intervention (if negative outcomes were generated with high levels of the FOI; O'Donnell, 2008).

Second, many FOI measures are designed to serve annual evaluation and monitoring within a non-research context. One critical component of SWPBIS is the data driven decision making as to student performance and the optimal solutions to improve student outcomes via an effective and efficient implementation process (Algozzine et al., 2010; Horner & Sugai, 2015). Out of multiple sources of data, fidelity data provides functional information regarding action planning before and during implementation. Specifically, school teams may monitor progress on FOI scores monthly or quarterly, typically with their coach, and revise specific strategies (i.e., student/staff supports, training, coaching) documented in implementation plans. Repeated progress monitoring of FOI data is a useful tool for overall improvement of FOI itself and outcomes (Resnick et al., 2005). Likewise, annual evaluation of fidelity is conducted to determine if the school has implemented SWPBIS with strong enough fidelity to affect student outcomes, which always guides development of the implementation plan for continuous improvement (Algozzine et al., 2010). In particular, one critical decision based on FOI data of SWPBIS is related to the MTSS framework. Implementation of the MTSS suggests that schools should document sustained strong fidelity of Tier 1 or Tier 2 before adding more intensive supports (Mitchell, Bruhn, & Lewis, 2016). Therefore, FOI data is needed for schools to decide whether they are fully prepared to implement the advanced tiers of supports in addition to Tier 1.

Within a large scale of implementation, district or state levels of authority use FOI data for their policy decisions (Barrett, Bradshaw, & Lewis-Palmer, 2008). Specifically, state-level or district-level leadership teams use FOI data to decide implementation planning and expansions. Regular reviews of FOI data allow leadership teams to identify and address school needs for technical supports or other resources, and offer practical aids (i.e., training, coaching, funding, or coordinating other local events) to schools. Also, annual evaluation of FOI data promotes the local endeavor to expand the implementation of SWPBIS to more schools and guides implementation plans to build support systems enough to address increasingly diversified school needs along with local progress from initial stages toward sustainability of SWPBIS (Barrett et al., 2008).

Measurement of the FOI of SWPBIS. Technical rigor of FOI measures is essential for any formal assessment of an intervention (Schoenwald et al., 2011). Within research situations, FOI measures with the most robust validity are chosen to evaluate intervention efficacy or effectiveness. Methodologically, external observation is considered the most rigorous than other sources such as intervention providers (e.g., school teams, school staff in SWPBIS) or recipients/participants (e.g., students in SWPBIS). However, observation methods require considerable time and expense associated with training observers, development of scoring protocols, data collection, and scoring processes (Schoenwald et al., 2011). In particular, observation is less feasible for school-wide practices, or complex and large scales of intervention involved by multiple providers and locations (Ruiz-Primo, 2005). Therefore, FOI measures for research adopt observation-based checklists integrating both direct and indirect assessment strategies (e.g., completion of checklists after interviews, walk through observations, and reviews of products) and depend on experts or coaches' judgement based on clearly specified criteria. Specifically, FOI measures such as the School-Wide Evaluation Tool (SET; Sugai, Lewis-Palmer, Todd, & Horner, 2001) for Tier 1, and the Individual Student Systems Evaluation Tool (ISSET; Lewis-Palmer, Todd, Horner, Sugai, & Sampson, 2003) for Tiers 2 and 3 usually depends on external coach or experts' observations and interviews at considerable expense (Algozzine et al., 2010).

FOI measures for annual evaluation or progress monitoring are intended to capture the same construct as the research measures but in more efficient and feasible manners (Algozzine et al., 2010). Practitioner report methods are generally used via self-assessment tools with an emphasis on the accountability for the valued outcomes in routine practice (Schoenwald et al., 2011). Despite the usability of FOI measures in routine practice, self-assessment scores might be more susceptible to inflation (McIntosh et al., 2017; Noell et al., 2005; Wickstrom, Jones, LaFleur, & Witt, 1996). To gain more objective and precise assessment, some FOI measures are completed by school teams with a guidance of an external coach, who facilitates scoring and interpretation process, and clearly defined scoring rubrics to help consistent and accurate administration of self-assessment. For example, annual evaluation tools such as the School-Wide Benchmarks of Quality (BoQ; Kincaid, Childs, & George, 2005) for Tier 1, and/or Benchmarks of Advanced Tiers (BAT; Anderson et al., 2012), or Monitoring Advanced Tiers Tool (MATT; Horner, Sampson, Anderson, Todd, & Eliason, 2013) for Tier 2 and 3 allow school teams to conduct a self-assessment with external coaches' guidance once a year. Similarly, the Self-Assessment Survey (SAS; Sugai, Horner, & Todd, 2000) is used for annual evaluation but with an emphasis on needs assessment and staff feedback. Also, FOI measures for progress monitoring include the Team Implementation Checklist (TIC; Sugai, Horner, & Lewis-Palmer, 2001) for Tier 1 and MATT for Tiers 2 and 3. Both measures involve more concise structures and smaller numbers of items in comparison to different measures. Particularly, the TIC can be rated monthly or quarterly by school teams in 15 to 20 minutes.

Operationalizing FOI in SWPBIS. With some variations due to the nature of intervention, the concept of fidelity has been generally extended beyond intervention delivery to interventionist competence, and intervention recipients' responsiveness (Schoenwald et al., 2011). Consideration of intervention providers and recipients allows FOI measures to capture overall implementation process involved to multiple stakeholders, which is suited for the main purpose of using FOI in SWPBIS pertaining to implementation planning. The SWPBIS Evaluation Blueprint (Algozzine et al., 2010) defined FOI as "how faithfully the program was implemented relative to its original design and focus and the resources that were directed to it" (p. 15). This definition recognized the importance of capacity building for schools and districts as an intervention provider in a sense that SWPBIS operates based on available behavioral expertise and resources at schools or broader levels such as training, coaching, or other professional development. Also, Algozzine et al. (2010) noted that FOI data be used as "evidence that core features are in place" (p. 15), which implies adherence to core features of SWPBIS as the central definition of FOI. Core features of SWPBIS are commonly understood as operational features of behavioral supports that conforms to theoretical principles to change behaviors, and also those of systems supports needed to plan, coordinate, and manage implementation of school-wide behavioral supports. Not all of those features are essential, but several moderating or mediating elements have been shown to affect implementation and outcomes. Determination of what features are critical or supplemental may change over time as the definition of SWPBIS evolves in correspondence to increasingly accumulated research (Century et al., 2010).

Specification of FOI of SWPBIS can begin with operationalizing each core feature. Based on the existing consensus on FOI as a multi-faceted construct, Table 1 proposes conceptual framework of FOI of SWPBIS via application of the revised Ruiz-Primo (2005)'s analytic tool to the TFI Tier 1 Scale. Except differentiation which is not applicable to this study, four aspects including exposure, quality, competence and responsiveness were examined. However, a few revisions were made from Dane & Schneider (1998), and Ruiz-Primo (2005) because this study attempted to differentiate qualitative descriptions of service delivery (i.e., quality) from those of service providers (i.e., competence).

Table 1.

Mapping FOI of SWPBIS for TFI Tier 1

Intended									Enacted (Adherence)		
Intervention Dimensions		Intervention Components		Indicators (Items)	Aspects of fidelity				Low → High		
					Exposure(doses)	Quality (form/manners)	Competence	Responsiveness	0 Not implemented	1 Partially implemented	2 Fully implemented
Differentiation (TFI)	Tier 1	Implementation of core features	Expectations	- Behavioral expectations - Expectation defined - 5 or fewer - positive - Matrix posted - 90% staff can list 2/3 expectations	√	√	√				
			Teaching	- Teaching expectations - Formal system - w/ schedules - directly taught - across settings - 70% of students can list 2/3 expectations		√		√			
			Reinforcement	- Feedback & Acknowledgement - A formal system - by 90% + of staff samples - to 50% + of student samples	√	√	√	√			
			Consequence	- Problem behavior definitions - Clear definitions - Clear procedures for office-/staff-managed problems - Documented/Trained/ Shared w/ families		√		√			
				- Discipline policies - School policies/ procedures - Proactive, instructive or restorative - Consistently		√					
			Classrooms	- Classroom procedures - Tier 1 features within classroom - Formally & Consistent w/ school-wide system		√					

Table 1 continued

Intervention Dimensions		Intervention Components		Indicators (Items)	Exposure(doses)	Quality (form/manners)	Competence	Responsiveness	Not implemented (0)	Partially implemented (1)	Fully implemented (2)
					Aspects of fidelity				Enacted (Adherence) Low → High		
Differentiation (TFI)	Tier1	Support Systems	Professional development (PD)	- PD - Formal process - Orienting all staff on - Teaching - Acknowledging - Correcting - Requesting assistance		√	√				
			Teaming	- Team composition - Tier 1 team - Coordinator, administrator, a family member, and individuals w/ several competencies - Attendance 80% +	√	√	√				
				- Operating procedure - At least monthly - Meeting agenda - Minutes - Meeting roles - Action plan	√	√					
			Evaluation (data system)	- Discipline data - Graph/summary on frequency by behavior, location, time, and students - Fully accessible		√					
				- Data-based decision making - Discipline/academic outcome data - Monthly - Team Review/ "use" for decision making	√	√					
				- Fidelity data - Team review/use - Formally - Annually	√	√					

Table 1 continued

Intended									Enacted (Adherence)		
					Aspects of fidelity				Low → High		
									0	1	2
Intervention Dimensions		Intervention Components		Indicators (Items)	Exposure(doses)	Quality (form/manners)	Competence	Responsiveness	Not implemented	Partially implemented	Fully implemented
Differentiation (TFI)	Tier1	Support Systems	Evaluation (data system)	- Annual evaluation - Fidelity/outcome data - Annually - Shared with stakeholders in a usable format	√	√					
		Other Considerations (mediator/moderators)		- Student/family/ community involvement - At least 1/12 months - Input on Tier 1	√			√			
				- Faculty involvement - At least 1/12 months - Input on Tier 1	√		√				

Table 1 showed that a majority of features on the TFI were specified for quality aspects. A detailed qualitative description of intervention delivery for both structural (i.e., materials; e.g., lesson plans) and procedural (i.e., means; e.g., "directly" taught to students) aspects were provided. Quantitative aspects were relatively more detailed in support systems, such as teaming and evaluation, which need to be routinely carried out to guide implementation of positive behavior supports. In addition, the competence aspect addressed description of characteristics or qualifications of intervention providers (e.g., roles of team members), expected (minimum) levels of staff engagement in the intervention (e.g., 90% of staff can list 2/3 expectations, at least 90% of staff samples engage in acknowledgement systems, regular collection of faculty input), and contents of professional development supports. As to recipients' responsiveness, characteristics of students eligible for intervention (e.g., screening), student participation in intervention (e.g., student involvement) and student receipt of interventions (e.g., at least 70% of students can list at least 67% of the expectations) were mainly considered, whereas actual enactment is regarded as outcomes in SWPBIS. Overall FOI of SWPBIS can be comprehensively operationalized for relevant aspects in line with the prevention logic of change brought to bear on SWPBIS practices and systems. Thus, FOI of SWPBIS can be understood as the extent to which the core features of positive behavioral supports and systems are delivered to students within a continuum of needs by school faculty and staff trained to implement them in a comprehensive and consistent manner toward socially valued outcomes (adapted from Sanetti & Kratochwill, 2009).

Tiered Fidelity Inventory (TFI)

Intended Uses of the TFI. Although there are multiple FOI measures of SWPBIS tailored to each tier, there was no single tool to assess FOI at all three tiers prior to the each tier, there was no single tool to assess FOI at all three tiers prior to the TFI (Algozzine et al., 2014; McIntosh et al., 2017). The TFI was developed as a complete index of FOI at all three tiers to establish core features of SWPBIS practices and systems; thereby, it allows school teams, decision-makers, evaluators and researchers to comprehensively and efficiently measure FOI at any tier in a common format and language across various types of schools (Algozzine et al., 2014). The TFI consists of three scales: Tier 1 (universal supports), Tier 2 (targeted supports), and Tier 3 (intensive supports). Each scale can be separately or jointly administered, which promotes different uses as: (a) an initial assessment of the current status of implementation, or a need identification pertaining to SWPBIS; (b) a progress monitoring tool to guide action planning and improvement process for tiers of focus; (c) a formative annual evaluation tool for the tiers of which have been already in place; and (d) an index of sustained implementation (Algozzine et al., 2014; McIntosh et al., 2017). As a decision making tool, the TFI can serve different needs of schools at different levels of implementation.

Scoring of the TFI. The TFI is organized into three scales corresponding to each tier and 10 subscales (Algozzine et al., 2014). Specifically, Tier 1 involves 15 items representing essential features of universal support practices under 3 subscales including Teams (2 items), Implementation (10 items), and Evaluation (3 items). Tier 2 offers 13 items indicating features of targeted supports from 3 subscales, including Teams (4 items), Interventions (5 items), and Evaluation (4 items). Tier 3 assesses 17 core features of intensive supports in 4 subscales including Teams (4 items), Resources (3 items), Plans (6 items), and Evaluation (4 items). School teams, facilitated by an external coach, self-rate items on the TFI using a 3-point Likert-type scale (0 = *not implemented*, 1 = *partially implemented*, 2= *fully implemented*). A scoring rubric offers specific criteria and available data sources for scaling responses. It is technically a self-assessment tool completed by school teams, but teams are strongly recommended to work with an external coach as a facilitator. However, it should be noted that data be gathered prior to the team meeting for scoring the TFI. In particular, the TFI Walkthrough (used to administer Tier I subscale of the TFI) can be completed, in which a minimum of 10% of staff (or 5 at small sized schools) and at least 10 students are randomly interviewed regarding their recognition of expectations, and their experiences teaching or being taught expectations, and providing or receiving rewards recently (Algozzine et al., 2014). All materials are freely available at http://www.pbis.org. The TFI can be scored and entered online at http://www.pbisapps.org or using pencil and paper.

After scoring all items, subscales and total scale scores of Tiers 1, 2, and 3 are first generated by summing all obtained points from relevant items (Algozzine et al., 2014). These raw scores are transformed to the percentage of SWPBIS implementation through dividing awarded points by maximally possible points. Specifically, the percentage scores for each and/or all of three tiers on the TFI indicates the percentage of core features actually enacted at Tiers 1, 2, or 3 and at all three tiers. To assist school teams to determine whether the current FOI of SWPBIS is adequate for producing or sustaining effectiveness, a criterion score of 70% is proposed for Tier 1 (McIntosh et al., 2017). This cut-off was selected based on preliminary comparison to the existing measures of SWPBIS fidelity (the BoQ and the SET), which had shown that when the criterion level is met (at least for Tier 1) there is likely to be improvement in student outcomes (Cohen, Kincaid, & Childs, 2007; Horner et al., 2004). Also, subscale and item scores (as raw or percentage scores) are reported to inform action planning for school teams.

Validation Process of the TFI. To establish validity evidence to support the interpretation and uses of TFI scores in certain contexts, it is necessary to identify assumptions that are expected to be true if TFI scores faithfully reflect the construct to be measures (Chapelle et al., 2010). Table 2 presents a series of inferences and assumptions linking observed performances to the proposed score interpretations for decision making.

If the assumptions underlying each inference are proven true by empirical or theoretical evidence, they could offer backing for the relevant inference.

Table 2.

Inferences and Assumptions on TFI Score Interpretation and Use

Inference		Interpretive arguments
From the target domain to the data: Collections of data from the proposed sources (e.g., walkthrough observation, documents, or interviews) corresponding to indicators on the TFI reveal the current status of implementation of core features of SWPBIS at schools.		
Construct & Domain description	Assumptions	− Core features of SWPBIS practices and systems can be identified for important aspects of the FOI. − The items on the TFI are reflective of the core features of SWPBIS implementation. − Data collection pertaining to item indicators from the proposed or equivalent data sources on the TFI are applicable at schools.
	Evidence	− The TFI was designed as a comprehensive and feasible measure of the FOI of SWPBIS based on the accumulated literatures, the existing instruments (e.g., BoQ, SET, ISSET) and unpublished measures (Algozzine et al., 2014; McIntosh et al., 2017). − Development of instruments including contents and scoring criteria were based on a systematic process to which experts and local users involved, and the follow-up revisions (McIntosh et al., 2017). − The expert panel (and also school teams and coaches at the following pilot tests) reported that the TFI contents and structures, and scoring criteria are valid for measuring important aspects of the FOI of SWPBIS (McIntosh et al., 2017).
From the data to the observed scores: The assembled data is evaluated by the school team using the scoring rubric to accurately and consistently generate the scores that reflect the FOI of SWPBIS at three tiers.		
Scoring	Assumptions	− The scoring rubric is clear and specific enough to guide evaluation of the FOI of SWPBIS. − The proposed scoring process minimize the potential influence from any inflation due to bias in self-assessment. − The scoring rubric is applied consistently across different conditions.
	Evidence	− From the initial pilot tests, school teams and their coaches reported that the TFI was easy and straightforward to administer, score, and interpret (McIntosh et al., 2017). − The pilot study found consistent applications of the scoring criteria across raters (the coach alone vs the coach with the school team), and over two weeks (McIntosh et al., 2017).

Table 2 continued

Inference		Interpretive arguments
From the observed scores to the universe scores (in line with the construct): The observed scores on the TFI are comparably invariant over replications to be estimated as the universe scores that are attributed to the construct of the FOI of SWPBIS.		
Generalization	Assumptions	– A sufficient number of features of SWPBIS practices and systems are specified on the TFI to deliver stable estimates of school-level FOI of SWPBIS. – The internal structure of the TFI is consistent with theoretical attributes underlying the construct of the FOI of SWPBIS. – TFI scores at item-, subscale-, and scale-levels are relatively invariant across fairly broad conditions of the measurement. – School performance on the TFI varies depending on years of implementing SWPBIS and/or the amount and quality of state- or district-level resources (e.g., initiatives, funding, technical support).
	Evidence	– The large-scale validation study found strong internal consistencies respectively for three scales of the TFI, and for all scales as a whole (McIntosh et al., 2017). – Both the 3-factor and 10-factor model displayed strong fit to the data, indicating that the TFI related item scores to underlying construct of the FOI of SWPBIS (Massar et al., in preparation).
From the universe scores to the target domain (outside the measure): The universe scores on the TFI are associated with the other equivalent or relevant performances or scores from different measures as explained by underlying theory and/or research within a target domain of the FOI of SWPBIS.		
Extrapolation	Assumptions	– Universe scores (latent traits) of each tier co-vary with those on different measures of the FOI of SWPBIS at corresponding tiers. – Universe scores (latent traits) on the TFI are predictive of the improvements in the intended outcomes such as staff perceptions or behaviors, students' academic and behavioral performance, and school culture (e.g., climates, safety, organizational health).
	Evidence	– The scores of each tier and all tiers as a whole on TFI were significantly correlated with different measures (BoQ, SAS, TIC, BAT Tier II, III, and total) of corresponding tiers. Specifically, correlations were stronger when teams with external coach scored two measures than when teams without external coach did (McIntosh et al., 2017).
From the score interpretation to the score use: The proposed interpretation of the TFI scores are useful for ongoing monitoring and evaluation to determine the current status, develop and revise action plans, and guide the implementation of SWPBIS with sustainable fidelity.		
Utilization (Decision)	Assumptions	– The meaning of TFI scores is clearly and accurately interpreted by school teams, external coach, and district/state-level staff. – Use of the TFI helps school teams develop an action plan and implement SWPBIS with adequate levels of the FOI.

Table 2 continued

Inference		Interpretive arguments
Utilization (Decision)	Assumptions	– Use of the TFI helps school teams complete a full multi-tiered support system in effective and efficient manners. – If the cut-score as an indicator of adequate-level of the FOI of SWPBIS is recommended to assist school teams for their decision making, the properties of the cut-score support the implications associated with the trait label; and the use of the cut-score provide a positive impact on actual stakeholders' decision making process.
	Evidence	– From the initial pilot tests, school teams and their coaches reported that the TFI was easy and straightforward to administer, score, and interpret (McIntosh et al., 2017).

Domain Description. First, data collections from the proposed sources (i.e., walk through observation, interview, review of documents) corresponding to specified indicators on the TFI are expected to reflect the target domain that the TFI intends to measure. To support this use, three assumptions can be evaluated by underlying theory, experts and/or other research as follows: (a) core features of SWPBIS practices and systems can be specified for important aspects of FOI, (b) the items on the TFI represent the core features of SWPBIS implementation, and (c) data collections pertaining to items from the proposed or equivalent sources on the TFI are applicable at schools.

Relevant to these assumptions, McIntosh et al. (2017) studied content validity regarding the TFI. Results indicated a systematic process to develop and refine the instrument involving experts and a pilot test with school users. For instance, the TFI was developed based on a compilation of the existing SWPBIS fidelity measures (i.e., TIC, SAS, BoQ, SET, MATT, BAT, ISSET), and also unpublished measures used in Florida, Illinois, Maryland, Missouri, and North Carolina (Algozzine et al., 2014; McIntosh et al., 2017). To guide a refinement of the TFI, 12 experts, including experienced researchers and school- and district-level implementers, were invited to evaluate content validity of the TFI. For each item, experts were asked to use a 4-point scale to rate (a) the extent to which individual items indicated important aspects of FOI (item validity), (b) the extent to which individual items were relevant to the proposed subscale (factor structure), and (c) the extent to which the scoring criteria were appropriate (scoring). They were also asked to determine whether each item should be retained or not. Then, evaluation of the measure as a whole (e.g., directions, response format, overall content validity) and specific feedback for revision were delivered by the panel. As a result, the expert reports with strong agreements indicates the overall mean score of Content Validity Index was .92 (Tier 1 = .95, Tier 2 = .93, Tier 3 = .91), which means that 92% of experts reported high levels of agreements (i.e., *strongly agree* or *agree*). As for the item-level responses, only a few items in each scale were rated below the criterion (.80) due to lack of relevance to critical features, clarity, or common language, which led to the revision process in item description and scoring criteria. This research indicates a carefully designed development process via both theory-based design and expert review, which offered a logical link from contents of the TFI to the target domain of FOI of SWPBIS.

In sequence, the pilot research (McIntosh et al., 2017) with ongoing revision was conducted from school teams and their coaches at 15 schools across 5 districts recruited by the state leadership teams in Connecticut, Michigan, Missouri, North Carolina, and Oregon. From the usability survey, they highly agreed (process = 100%; scoring = 93%; validity = 100%) that the TFI was easy and straightforward to administer, score, and interpret. Noteworthy is that the follow-up revisions after two studies were made based on feedbacks from experts and actual users to enhance clarity, relevance, and common language use. Therefore, research supported that data collection and review based on the TFI were feasible at school settings. Overall, research provided evidence to support the underlying assumptions of inference from the target domain to data collection.

Scoring. Evaluation of the assembled data using the scoring rubric is expected to generate raw scores in consistent and accurate manners across raters, occasions, and measurement conditions. Scoring inference implies three assumptions: (a) the scoring rubric is clear and specific enough to guide evaluation of the FOI of SWPBIS, (b) the proposed scoring process minimizes the potential influence from implicit bias of self-assessment, and (c) the scoring rubric is applied consistently across different conditions of measurement. As earlier, the pilot test (McIntosh et al., 2017) supported that the TFI, including the scoring rubric, was easy and straightforward to use. From the same pilot study, the test-retest reliability with a two-week span was examined, revealing a consistent application over time (intra-class correlation [ICC] = .99). Furthermore, interrater reliability across each tier, and all items was examined through comparing the scores obtained from the coach's sole administration and those from the coach working with the school team, which revealed high reliability (ICC = .99 for all). Although there still remains the need to explain whether scores by school teams and their coach together are invariant from those by school team alone, this study offered an evidence for accurate application of the proposed scoring process in which school teams are encouraged to work with external coach when administering the TFI (Algozzine et al., 2014). The current evidence indicates that the scoring rubric is clear and understandable for school teams and their experts; administration by school teams working with their coach generated consistent scores in accordance to the detailed scoring criteria. It can be concluded that research supports the scoring inference extending data collections to scores.

Generalization. To infer a school's observed scores on the TFI of a particular time point of the measurement (as a sample) to more broadly expected scores, or universe scores that the school is highly likely to receive under all possible conditions of the measurement (e.g., items, locations, occasions, raters), it is necessary to justify that the proposed measurement procedure draws a representative sample (or observed score) from the universe of observations under the target domain of FOI of SWPBIS (Kane, 2006). Thus, the generalization inference requires the evaluation of the following assumptions: (a) a sufficient number of items are specified on the TFI to deliver stable estimates of school-level FOI; (b) internal structure of TFI scores are consistent with theoretical attributes of the construct of FOI of SWPBIS; (c) TFI scores at item-, subscale-, and scale-levels are likely to be invariant across fairly broad conditions of the measurement; d) school performance on the TFI may vary depending on years of implementing SWPBIS and/or the amount and quality of state-or district-level resources.

One basic assumption for generalization is that individual items (as a facet of the measurement) on the TFI are representative for universe of all possible items. Specifically, a set of items intended to measure the same construct on a same scale are expected to be large enough to reduce the sampling variability, and to be interrelated (Chapelle et al., 2010; Kane, 2006). One source of evidence is available in McIntosh et al (2017)'s large-scale validation study, in which 789 schools across seven states during the 2013-14 school year were involved. Internal consistency via coefficient alpha, which is a function of number of items, the variance of and covariance across individual items

(Thorndike & Thorndike-Christ, 2010), was calculated, resulting in strong internal consistency (all = .96; Tier 1 = .87; Tier 2 = .96; Tier 3 = .98). This result indicates that observed scores of TFI items prominently and coherently captured the universe of scores under the target domain of FOI of SWPBIS.

Another assumption pertains to relations between universe scores and the underlying theory of the construct. Although the initial evidence of content validity of the TFI (McIntosh et al., 2017) makes plausible that universe scores obtained from observing and scoring the target domain relate to a construct of FOI of SWPBIS, this assumption under the generalization inference needs to be justified by empirical evidence on whether the factorial structure on TFI scores are in line with theoretical attributions underlying the construct of FOI of SWPBIS; the alignment of model structure of the TFI to the underlying theory can be generalized to broad levels of conditions (Cubilo, 2014; Kane, 1992). The relevant evidence can be found in Massar et al. (in press). They conducted a confirmatory factor analysis based on the scale (3-factor model) and subscale (10-factor model) structures of the TFI from a large-scale sample of 1,708 schools across 25 states (including a larger proportion of elementary and middle schools and schools in California and Illinois), and found that variances in item scores on the TFI were significantly related to those in their corresponding factors under not only the 3-factor model but also the 10-factor model with strong fit to the data. This result indicated that the TFI scale- and subscale structures related item scores to scale- or subscale level universe scores in line with the underlying construct of FOI of SWPBIS. Based on a fairly large sample size, both factorial models (align with scale and subscale structure) can be extended to the universe score for the larger settings given the demographic portfolio of the sample.

Regarding the generalization inference, the existing research supports the argument that TFI items coherently measure the theoretical constructs of FOI of SWPBIS; TFI subscale and scale scores can be extended to universe scores for broader types of schools (but conservatively limited to elementary and middle schools in California and Illinois). However, the latter two assumptions (i.e., c and d) still need more evidence via ongoing research. Typically, generalization draws upon statistical evidence on the amount and sources of measurement errors to demarcate a trustworthy portion of universe scores (Cubilo, 2014; Kane, 2006). In recognition that variances in TFI scores are affected by school contexts, raters, and occasions (time), there remain some needs for reliability and generalizability studies that aim to identify sources of variance, in particular, construct-irrelevant but systematic errors influencing a measurement. For example, the factorial model validated by Massar et al. (in press) can be replicated with an emphasis on group invariance to answer whether unstandardized factor loading(s) and/or intercepts of corresponding items on the TFI are invariant across school contexts (e.g., grade, type), raters (e.g., with versus without an external coach), and occasions of measurement. To gain more evidence on generalization of item-level scores, differential item functioning can be examined to test whether different conditions of measurement produce a gap in probability of success on the item given the same level of latent traits. Also, TFI scores are inherently time-variant data. With relation to stability of between school variances from one occasion to the next (largely on a year basis), it is necessary to examine longitudinal patterns of school performances on the latent traits (factorial scores) from the TFI.

Extrapolation. The universe scores in line with the construct of the interest are expected to have a connection to target scores, outside the TFI, measuring other relevant or equivalent constructs (Kane, 1992). The extrapolation inference can be evaluated with the following assumptions: a) universe scores of each tier (extended from observed scores on the TFI) co-vary with those on different measures of FOI of SWPBIS at corresponding tiers; b) universe scores on the TFI predict the improvements in the intended outcomes such as staff perceptions or behaviors, students' academic and behavioral performance, and school culture (e.g., climates, safety, organizational health). Relevant to these assumptions, McIntosh et al. (2017) conducted a concurrent study as a part of large-scale validation study and found that the scores of each tier and all tiers as a whole on the TFI were significantly correlated with those from different measures (BoQ, SAS, TIC, BAT Tier II, III, and total) of corresponding tiers. Moreover, correlations were compared across scoring conditions, finding that scores of two measures become correlated more strongly when teams worked with an external coach than without (McIntosh et al., 2017). This research extended universe scores to equivalent target scores measuring the same construct of FOI of SWPBIS and also provided more information regarding potential effects of the coach for administration of the TFI. However, additional studies are needed to justify the second assumption regarding connection to student outcomes, which is the focus of the current study.

Utilization. If the proposed score interpretation is supported based on the justification of logically chained inferences from the target domain to extrapolation, it is necessary to connect the proposed score interpretation to school teams' actual use of TFI scores for decision-making regarding implementation of SWPBIS. Utilization inference can be evaluated based on the following assumptions: (a) the meaning of TFI scores is clearly and accurately interpreted by school teams, external coach, and district- or state-level staff; (b) the use of the TFI helps school teams develop an action plan and implement SWPBIS with adequate FOI; (c) the use of the TFI helps school teams implement a full MTSS model in effective and efficient manners. As indicated by McIntosh et al. (2017), the pilot study supported that the TFI was easy and straightforward for school teams and their coaches to interpret. Despite their satisfaction with using the TFI as a single tool to assess the three tiers, school teams and coaches reported that that Tier 2 and Tier 3 scales could replace other measures whereas Tier 1 scale might need supplement from other Tier 1 FOI measures for different purposes (i.e., the TIC for initial implementation, BoQ for in-depth examination, SAS for gaining staff input). With relation to specific purposes (e.g., the initial assessment, monitoring, annual evaluation), further research is needed to focus on whether the use of the TFI is helpful and cost-effective for school teams for decision making; whether the use of the TFI has a positive impact on establishing SWPBIS systems and practices in their schools. To guide decision making using the TFI scores, the cut-score of 70% is recommended as an indicator of adequate-level of FOI of SWPBIS, but there was no research to rationale the use of cut-score on the TFI. Thus, future research is needed to provide evidence to support the following assumptions: (d) the properties of the cut -score support the implications associated with the trait label and (e) the use of the cut-score provide a positive impact on actual stakeholders' decision-making process.

Summary on the accumulated validity evidence for the TFI. TFI percentage scores at each tier are interpreted as the proportion of enacted core features of SWPBIS practices and systems across tiers (McIntosh et al., 2017) as an indicator of FOI of SWPBIS. Individual items and subscale scores interpreted as an indicator of the extent to which individual components or features of SWPBIS are in place. The current validity evidence from two validation studies (Massar et al., in press; McIntosh et al., 2017) suggested that the proposed score interpretation be valid; and the TFI is feasible and comprehensible for school teams and their coaches to use. However, there still remain loosely connected arguments in need of coherent logical links of chained inferences. Particularly regarding the extrapolation inference, there is no evidence on theoretically explained relations between the TFI scores and the intended outcomes of SWPBIS. Extended from the existing evidence on internal structure of SWPBIS, the current study will replicate the measurement model of the TFI to make the generalization inference more plausible and also warrant the theory-based extrapolation inference via connecting the universe score of the TFI to the outcomes of SWPBIS.

Tiered Fidelity Inventory and Predicted Outcomes

This study will investigate the predictive validity of the TFI with an emphasis on student and staff behavior outcomes. Multi-tiered supports require school teams to clearly define and regularly monitor either academic or behavioral outcomes for students at each tier. Thus, the SWIS Suite, a web-based data management system for schools implementing SWPBIS is tailored to MTSS, involving School-Wide Information System (SWIS) for universal supports, Check-In Check-Out (CICO-SWIS) for CICO, one of targeted supports, and Individual Student Information System (ISIS-SWIS) for individualized student supports (PBISApps, 2016a, September).

Accumulative effects of MTSS on student outcomes. However, distinctive management of student outcome data across three tiers does not mean that each tier has separate effects because delivery of either Tier 2 or 3 supports does not mean the removal of individual students from less intensive supports (Lane et al., 2013; Horner & Sugai, 2015). Rather, advanced tiers supplement Tier 1 prevention efforts for students in need of additional intensive supports. Likewise, the effects of fidelity of each tier cannot be separate from but are interconnected to those of other tiers. Specifically, a strong Tier 1 system allows schools to enhance system-level readiness for additional tiers (Fairbanks, Simonsen, & Sugai, 2008; Fairbanks, Sugai, Guardino, & Lathrop, 2007). Thereby, this study will investigate the predictive validity based on the assumption that all student outcome data at each of three tiers is more strongly related to fidelity at the corresponding tiers.

Tier 1 student behavior outcomes data. One readily accessible measure of problem behavior at schools is exclusionary school discipline, including ODRs. Many studies (Irvin, Tobin, Sprague, Sugai, & Vincent, 2004; Irvin et al., 2006; McIntosh, Campbell, Carter, & Zumbo, 2009; Pas, Bradshaw, & Mitchell, 2011) indicate that school discipline outcomes are valid and reliable for interpreting Tier 1 patterns of problem behaviors. Research indicates effectiveness of SWPBIS implementation for reducing school discipline outcomes (Bradshaw et al., 2010; Horner et al., 2009). In particular, FOI of Tier 1 was shown as a significant predictor of decreased problem behaviors. Therefore, it can be hypothesized that schools with higher Tier 1 fidelity are likely to have lower rates of ODRs.

In addition, prior studies (McIntosh, Campbell, Carter, & Zumbo, 2009; Walker, Cheney, Stage, Blum, & Horner, 2005) found that students with 2 or more ODRs were significantly more likely to have higher scores on standardized ratings of problem behaviors (e.g., Problem Behavior Scale of the *Social Skills Rating System*). In particular, McIntosh and colleagues found that ODR cut points such as 0 to 1, 2 to 5, and 6 or more put students respectively at "on-average," "at-risk," and "clinically significant" classifications based on the scores of the Externalizing Composite of the *Behavior Assessment Scale for Children 2* (Reynolds & Kamphaus, 2004). In accordance, it can be hypothesized that schools with high FOI of Tier 1 systems would have a larger proportion of students with 0 to 1 ODRs than those with more than 1 ODRs; and a lower proportion of those with 6 or more ODRs than those with less than 6 ODRs.

Boneshefski and Runge (2014) noted that the benefits of SWPBIS might not be even for all subgroups of students. Vincent, Swain-Bradway, Tobin, and May (2011) revealed from a national study with schools implementing SWPBIS that African American students still displayed the highest rates of ODRs despite overall reductions for both general and minority groups over three years of SWPBIS implementation. However, Vincent et al. (2011) noted that schools implementing SWPBIS with adequate FOI showed significantly smaller discipline disproportionality than those without adequate FOI. As such, Boneshefski and Runge (2014) proposed that SWPBIS framework might provide a useful context for schools to regularly disaggregate school discipline data and examine disproportionality, which offers a question of whether and to what extent adherence to FOI of SWPBIS is associated with decreased discipline disproportionality.

Recently, federal regulation (Determining Significant Disproportionality, 2016**)** recommended using risk ratio (dividing the risk index for a particular racial or ethnic group within the district by the index for all other students within the district) or alternate risk ratio (dividing the risk indexes for a particular racial or ethnic group within the district by those for the rest of groups in the state due to insufficient size of analyzed groups enough to generate the stable risk ratio) with the reasonable risk ratio thresholds and the minimum size of racial/ethnic groups analyzed. Despite the easy interpretability (Boneshefski & Runge, 2014, Parrish, 2002), the risk ratio exhibits instability, based on small variations in size of either group compared, and cannot be calculated for schools with zero risk index for the reference group as the denominator due to missing referral risk ratios for comparison groups(Girvan, McIntosh, & Smolkowski, 2018). To compensate for such limitations, another comparative metric of racial difference of risk indices for the referrals between two groups, with relatively stronger stability (Girvan et al., 2018), can be used. It can be hypothesized that schools with high fidelity of SWPBIS implementation may use their data to identify and respond to disproportionate outcomes in terms of race (in particular, African American students), and thereby have comparatively lower referral risk difference.

Tier 1 student academic outcomes data. Effective implementation of SWPBIS can provides students with a positive, safe, and predictable environment that is required for effective learning. Horner et al. (2009) found significant differences in the proportion of third grade students satisfying state-wide standards in reading achievement tests between pre- (Time1) and post-training (Time2) condition for the treatment group who received training after Time 1, and between the treatment and control/delay group at

Time 2 before control/delay group received training. In addition, Gage, Leite, Childs, and Kincaid (2017) found that implementation of SWPBIS with FOI meeting or exceeding the criterion on BoQ was significantly (but weakly) associated with higher proportions of students satisfying the state-level standards in reading and math. In accordance, it can be hypothesized that schools with high fidelity of SWPBIS implementation would have higher proportions of students meeting or exceeding expectations on state reading and mathematics achievement tests.

Tier 2 student behavior outcomes data. CICO is one of the widely used Tier 2 behavior support interventions (Hawken, Bundock, Barrett, Eber, Breen, & Phillips, 2015; Maggin, Zurheide, Pickett, & Baillie, 2015). CICO intends to provide students with frequent and structured opportunities for positive consequences for their appropriate behaviors aligned with school-wide expectations. Tier 2 interventions consist of ongoing multiple interventions readily accessible to eligible students within schools (McIntosh et al., 2009). To implement Tier 2 interventions in a consistent and timely manner, school teams need to define clear decision rules to identify students in needs of additional supports, match them to appropriate interventions, determine student progress, or to conclude exit from secondary interventions (Lane et al., 2013; Algozzine et al., 2014). Schools with a solid multi-tiered system should be able to support approximately 10 to 15% of students through Tier 2 supports. Although Crone, Hawken, and Horner (2010) proposed from their manual that one CICO coordinator support 15 to 20 elementary school students or 20 to 30 secondary school students, this size could be only small proportions (e.g., around 3 to 4% out of 800 students) for many schools (Hawken et al., 2015). Hawken et al. found from 54 schools that schools with greater than 70% scores on

Foundations and Targeted ISSET showed larger proportions of students on CICO (11.71%) than those with lower than 70% (6.57%). This result indicates that schools with well-developed capacity for Tier 2 can support a larger % of students at Tier 2 than schools with less capacity.

One component of CICO is a daily progress report (DPR), which students carry throughout a daily check-in check-out routine to receive structured feedback and earn points from teachers for their adherence to the listed behavioral expectations (Hawken et al., 2015). In particular, DPR points "earned" for the day are reviewed with students and used to deliver praise and feedback by the coordinator during the check out. Scott et al. (2010) noted that a daily point card would provide a sensitive and feasible tool to collect data on student progress at Tier 2. Many studies used DPR points as a measure of the extent to which students engage in expected behavior (Hawken et al., 2014).

Within the CICO routine, students earning at least 80% of possible points per day are usually considered successful on CICO or having no needs of additional supports (Crone et al., 2010; Hawken et al., 2015). Hawken et al. (2015) found from schools in the Illinois PBIS Network that greater than 80% of students tended to earn at least 80% DPR points. This result indicates that a majority of students followed behavioral expectations within a CICO context. As effective implementation of Tier 2 intervention promotes student adherence to school-wide expectations for students enough to exit Tier 2 supports, it can be hypothesized that schools with higher Tier 2 FOI would have a larger proportion of students with at least 80% DPR points per day on average.

Although DPR points earned are useful indicators of student progress on CICO, the intervention is ultimately intended to change school discipline records that might offer evidence for the school team to discontinue the intervention. Hawken et al. (2014) found from their literature reviews that many studies demonstrated positive effects of CICO on reductions in ODRs. Schools with strong capacity to implement Tier 2 interventions are expected to improve student disciplinary outcomes. Therefore, it can be hypothesized that schools with higher Tier 2 FOI will have lower levels of ODRs per 100 days per student.

Tier 3 student behavior outcomes data. Tier 3 supports offer the highest intensity of behavioral supports for students whose behaviors do not respond to Tier 1 or 2 supports. Tier 3 supports involve individualized assessments and support plans, student support teams, and data-driven decision making (Horner & Sugai, 2015; Scott, Anderson, & Spaulding, 2008). The most widely-used Tier 3 interventions are FBA (functional behavior assessment) – based interventions. An FBA involves formal and informal assessment to identify function of problem behaviors within the surrounding context and to develop function-based behavior support plans. The FBA process operates with flexibility based on increasingly formal, individualized, and comprehensive processes (Scott et al., 2010).

FBA-based interventions have been shown effective for reducing problem behaviors (Gage, Lewis, & Stichter, 2012; Lane et al., 2007; Miller & Lee, 2013). For example, Lane and colleagues (2007) found from single subject research a functional relation between function-based interventions and reductions in problem behaviors for two students, respectively each at an elementary and a middle school, who had not been responsive for Tier 1 and 2 interventions. Despite few studies directly investigating relations between FBA based interventions and school disciplines, strong evidence on effectiveness for reducing problem behaviors at school settings offers the potential hypothesis that schools implementing Tier 3 interventions with fidelity might lead to low rates of school discipline events.

Data Use Activity. One key component of SWPBIS is data collection and use for decision making (Horner et al., 2005). Regular review of student data allows school teams to track student progress, learn schoolwide patterns of student behaviors, and identify current needs of supports, which offers iterative feedback to school teams for their decisions on implementation process. In particular, McIntosh et al. (2013) found that a school team's data use and sharing for decision making were associated with sustained, adequate FOI. In this regard, this study will examine school team data use, with a hypothesis that schools with strong FOI of SWPBIS are likely to engage in frequent review of student outcome data, as well as update action plans and implementation strategies.

Consideration of Years of Implementation. FOI is a time-varying construct that interact with contextual factors within and/or surrounding school buildings (McIntosh, Mercer, Nese, & Ghemraoui, 2016; Turri, Mercer, McIntosh, Nese, & Strickland-Cohen, 2016). Multiple years of implementation efforts are often needed to reach adequate levels of fidelity with varying speeds, and numerous schools tend to fail to sustain fidelity or abandon the initiative before meeting the criterion of fidelity (McIntosh et al., 2016). Research (Turri et al., 2016) found that higher levels of FOI of SWPBIS and lower levels of perceived barriers to implementation were found in schools from the full operation group and sustainability group than in those from initial implementation group. McIntosh et al. (2016) noted that the typical 1 to 3 year period of initial implementation might contain numerous risk factors for abandonment but relatively stronger impacts of fidelity on student outcomes within a fast changing environment. In this respect, it will be hypothesized that schools with longer years of implementing SWPBIS have higher TFI scores (Childs, Kincaid, George, & Gage, 2016; Turri et al., 2016); and fidelity effects vary based on years of implementing SWPBIS.

Consideration of School Demographic Variables. Much research (Freeman et al., 2016; Turri et al., 2016; Flannery et al., 2013) has examined school demographic variables as a predictor of fidelity of implementation and student outcomes. Based on the previous findings, this study hypothesized that elementary schools tend to have higher fidelity and academic outcomes, and lower levels of problem behaviors compared to middle or high schools (Childs et al., 2016; Freeman et al., 2016; Turri et al., 2016); schools with larger percentage of minority students or free-reduced lunch may be more likely to have lower levels of fidelity, academic and behavioral outcomes (Freeman et al., 2016; Flannery et al., 2013; Turri et al., 2016).

Research Questions

To validate the TFI in terms of its relation to student outcomes, this study will investigate whether TFI scores are associated with desired academic and behavioral outcomes. Specifically, the following questions will be answered:

1. To what extent are TFI Tier 1, 2, and 3 scores associated with (a) major ODRs per 100 students per day, (b) referral risk ratio for African American students, and (c) the proportion of students with 0 – 1 major ODRs in 2016-17 after controlling for student behavior outcomes of the prior year and other contextual variables?

2. To what extent are TFI Tier 1, 2, and 3 scores associated with proportions of students meeting or exceeding state-level standards in reading or math in 2015-16 after controlling for student academic outcomes of the prior year and other contextual variables?
3. To what extent are TFI Tier 1, 2, and 3 scores associated with (a) proportions of students enrolled in CICO, (b) students meeting 80% or more of their daily points, and (c) major ODRs per 100 students per day for students in CICO in 2016-17 after controlling for student behavior outcomes of the prior year and other contextual variables?
4. To what extent are TFI Tier 1, 2, and 3 scores associated with (a) years implementing SWPBIS, (b) counts of viewing SWIS Core or Additional Reports, (c) counts of viewing CICO Reports, (d) frequency of TFI administrations, and (e) number of fidelity measures used in 2016-17 for schools?

CHAPTER II

METHOD

Participants and Settings

This cross-sectional study focused on school-level data involving TFI scores and academic and/or behavioral outcome measures in the school years of 2014-15 to 2016-17. Depending on types of student outcome measures across research questions, different datasets and subsamples were used. First, the investigation of associations between the TFI and behavioral outcomes (in research question 1, 3, 4) used the sample of schools reporting the TFI in 2016-17 and school-wide discipline outcomes or behavioral outcomes of CICO intervention in 2015-16 and 2016-17 pulled from PBISApps, a web-based application to assist in data collections and data based decision making at schools (PBISApps, 2016a, 2017b). The second research question regarding associations between the TFI and academic outcomes focused on the sample of schools with the TFI and academic achievement data in reading and/or math pulled from state-level data centers in 2015–16.

Description of Sample

Question 1. The TFI prediction of school-wide behavior outcomes focused on the subsample of 1,691 elementary or secondary schools with both TFI Tier 1 scores in the 2016-17 school year and SWIS data consecutively in the 2015-16 and the 2016-17 school year. SWIS is a web-based data system designed to collect, summarize, and use school-wide discipline records for decision making, which is linked to different databases of the SWIS Suite (PBISApps, 2016a). These schools came from 34 states involving California (*n* = 287), Illinois (*n* = 182), Michigan (*n* = 136), Iowa (*n* = 87), Oregon (*n* =

124), Minnesota (*n* = 121), Wisconsin (*n* = 103), Washington (*n* = 96), Kentucky (*n* = 69), Vermont (*n* = 66), Missouri (*n* = 61), Connecticut (*n* = 37), Montana (*n* = 35), North Carolina (*n* = 33), Ohio (*n* = 33), New York (*n* = 32), Massachusetts (*n* = 27), Nebraska (*n* = 26), Mississippi (*n* = 21), Texas (*n* = 21), Arizona (*n* = 19), New Hampshire (*n* = 14), Pennsylvania (*n* = 13), Colorado (*n* = 10), Maine (*n* = 6), Tennessee *n* = 6), Virginia (*n* = 6), Indiana (*n* = 5), Maryland (*n* = 4), South Dakota (*n* = 4), Arkansas (*n* = 2), District of Columbia (*n* = 2), Nevada (*n* = 2), and Idaho (*n* = 1).

Those schools consisted of 1,218 elementary schools (72.03% out of 1,691 schools), 292 middle schools (17.27%), and 159 high schools (9.40%) as well as 22 other types of secondary schools (1.30%; e.g., 6 to 12th). Also, there were 1,609 regular schools (98.8% out of 1,628 schools), and 19 other types of schools (1.2%) such as special or alternative schools. Regarding the location, there were 476 schools (29.24% out of 1,628 schools) at urban areas, 544 schools (33.42%) at suburban areas, 252 schools (15.48%) at towns, and 356 schools (21.87%) at rural areas. With relation to enrollment, 138 schools (8.16% out of 1,691 schools) involved 200 or less students; 501 schools (29.63%) had 201 to 400 students; 563 schools (33.29%) had 401 to 600 students; 264 schools (15.61%) had 601 to 800 students; 225 schools (13.31%) had more than 800 students.

On average (out of 1,626 schools), the percent of students eligible for free or reduced-price lunch was 56.42% (*SD* = 24.75%; Range = 0 ~ 100%). Also, the percent of female ranged from 13.25 to 62.31 (*M* = 48.25%, *SD* = 3.14%). Related to racial proportion (out of 1,681 schools), the mean percent of Caucasian students was 55.35% (*SD* = 32.08%; Range = 0 ~ 100%), and that of African American was 9.96 (*SD* = 16.37%; Range = 0 ~ 99.49%). On average, 1.50% was Native American (*SD* = 7.76%;

Range = 0 ~ 98.99%), 5.08% was Asian students (*SD* = 9.69%; Range = 0 ~ 82.74%), 23.35% was Hispanic students (*SD* = 26.81%; Range = 0 ~ 100%), 0.35% was Hawaiian (*SD* = 0.98%; Range = 0 ~ 16.02%), and 4.40% was racially mixed (*SD* = 3.54%; Range = 0 ~ 27.27%).

With relation to experience of SWPBIS, those schools had implemented SWPBIS for 6.45 years on average (*SD* = 3.41; range = 1 to 18). The TFI T1 scores from the sample ranged from 6.67% to 100% (*M* = 82.55, *SD* = 15.90), and 83.7% met or exceeded 70% as the criterion of TFI T1.

Question 2. A total of 1,361 schools with both TFI Tier 1 scores in the 2015-16 school year and academic data (reading and/or math test scores) in 2014-15 and 2015-16 were used for the second research question. These time periods were selected due to limited accessibility to academic outcome data in 2016-17. These schools came from 16 states, including California (*n* = 317), Illinois (*n* = 201), Michigan (*n* = 123), Iowa (*n* = 113), Oregon (*n* = 112), Wisconsin (*n* = 89), Missouri (*n* = 85), Washington (*n* = 71), Minnesota (*n* = 64), North Carolina (*n* = 47), Kentucky (*n* = 42), Montana (*n* = 29), Ohio (*n* = 27), Texas (*n* = 23), North Hampshire (*n* = 11), and Virginia (*n* = 7).

The sample consisted of 962 elementary schools (70.68% out of 1,361), 277 middle schools (20.35%), and 111 high school (8.16%) as well as 11 other types (0.81%). Also, 98.97% (*n* = 1,347 out of 1,361) were regular schools, and the rest of schools (*n* = 14) were special or vocational schools. The schools located at urban areas (*n* = 403, 29.61%), at suburban areas (*n* = 468, 34.39%), at towns (*n* = 212, 15.58%), and at rural areas (*n* = 278, 20.43%). In terms of enrollment, 77 schools (5.66% out of 1,361 schools) involved 200 or less students; 390 schools (28.66%) had 201 to 400 students; 475 schools

(34.90%) had 401 to 600 students; 232 schools (17.05%) had 601 to 800 students; 187 schools (13.74%) had greater than 800 students.

Regarding student demographics, the proportion of female students ranged from 17.14% to 58.43% (*M* = 48.32%, *SD* = 2.92%). With relation to racial proportion, 54.26% (*SD* = 31.65%; range = 0 ~ 99.74%) was Caucasian; 25.28% was Hispanic (*SD* = 27.96%; range = 0 ~ 100%); 9.21% was African American (*SD* = 15.87%, range = 0 ~ 100%); 5.26% was Asian (*SD* = 10.02%, range = 0 ~ 84.33%); 1.31% was American Indian/Alaska Native (*SD* = 6.07%, range = 0 ~ 99.66); 0.37% was Hawaiian/Pacific Islander (*SD* = 0.94, range = 0 ~ 14.29); and 4.31% was from two or more races (*SD* = 3.41, range = 0 ~ 17.83). Also, the average proportion of students eligible for receiving free or low-priced lunch was 57.26% (*SD* = 23.74, range 0 ~ 100).

With relation to experience of SWPBIS, those schools had implemented SWPBIS for 5.89 years on average (*SD* = 3.46; range = 1 to 17). The TFI T1 scores from the sample ranged from 6.67% to 100% (*M* = 80.31, *SD* = 16.50), and 79.43% met or exceeded 70% as the criterion of TFI T1.

Question 3. The TFI prediction of behavioral outcomes for students enrolled in CICO involved 570 elementary and secondary schools having both TFI Tier 2 scores in the 2016-17 school year and CICO-SWIS data in 2015-16 and 2016-17. CICO-SWIS is a part of a host system of SWIS Suite, a web-based information system tailoring to schools implementing CICO (PBISApps, 2016b). There were 556 regular school (99.82% out of 557), and 1 special education schools (0.18%). Also, the subsample included 456 elementary schools (80.00% out of 570), 102 middle schools (17.89%), and 10 high schools (1.75%) as well as 2 secondary 6^{th} – 12^{th} schools (0.35%). Also, there were 151 schools (27.11% out of 557) at urban areas, 221 schools (39.68%) at suburban areas, 87 schools (15.62%) at towns, and 98 schools (17.59%) at rural areas. In terms of enrollment, 26 schools (4.56% out of 570 schools) included 200 or fewer students; 196 schools (34.39%) had 201 to 400 students; 195 schools (34.21%) had 401 to 600 students; 96 schools (16.84%) had 601 to 800 students; 57 schools (10%) had more than 800 students.

Regarding student demographics, the percent of female students ranged from 13.25% to 60.45% (*M* = 48.20%, *SD* = 2.97%). With relation to racial proportion, 55.83% (*SD* = 29.66%; range = 0 ~ 100%) was Caucasian; 23.41% was Hispanic (*SD* = 25.33%; range = 0 ~ 99.34%); 10.41% was African American (*SD* = 16.63%, range = 0 ~ 95.13%); 4.47% was Asian (*SD* = 7.04%, range = 0 ~ 63.17%); 1.00% was American Indian/Alaska Native (*SD* = 5.61%, range = 0 ~ 95.02); 0.40% was Hawaiian/Pacific Islander (*SD* = 1.03, range = 0 ~ 8.26); and 4.49% was from two or more races (*SD* = 3.24, range = 0 ~ 18.09). Also, the average proportion of students eligible for receiving free or low-priced lunch was 54.49% (*SD* = 24.18, range 0.45 ~ 100).

With relation to experience of SWPBIS or CICO, those schools had implemented SWPBIS for 7.68 years on average (*SD* = 3.42; range = 1 to 18), and CICO for 4.49 years on average (*SD* = 1.94; range = 2 to 8). The TFI T1 scores from the sample ranged from 13.33% to 100% (*M* = 87.55, *SD* = 12.19), and 89.18% met or exceeded 70% as the criterion of TFI T1. The TFI T2 scores ranged from 3.85% to 100% (*M* = 82.83, *SD* = 17.55), and 79.4% met or exceeded 70% as the criterion of TFI T2.

Question 4. The fourth research question regarding the concurrent validity of TFI scores used a subsample of 2,379 schools with TFI Tier 1 and/or Tier 2 scores and

SWIS and/or CICO-SWIS outcome data in the 2016-17 school year. Overall, 2,224 schools (98.06% out of 2,268) are regular school, and 44 schools (1.94%) were special, vocational, or alternative schools. Also, the sample included 1,695 elementary schools (71.22%), 398 middle schools (16.76%), and 222 high schools (9.33%) as well as 30 secondary (6^{th} to 12^{th}) schools (1.26%) and 34 all grade schools (1.43%). Regarding the location (n = 2,268), there were 689 schools (30.38%) at urban areas, 794 schools (35.01%) at suburban areas, 335 schools (14.77%) at towns, and 450 schools (19.84%) at rural areas. In terms of enrollment, 195 schools (8.21% out of 2,376 schools) enrolled 200 or fewer students; 676 schools (28.45%) had 201 to 400 students; 789 schools (33.21%) had 401 to 600 students; 397 schools (16.71%) had 601 to 800 students; 319 schools (13.43%) had more than 800 students.

Overall, student demographics showed almost equal proportion of female, M = 48.22%, SD = 3.55%, and Range = 11.11% ~ 100%. Regarding racial proportion, 52.39% (SD = 33.39%; range = 0 ~ 100%) was Caucasian; 27.14% was Hispanic (SD = 29.92%; range = 0 ~ 100%); 9.48% was African American (SD = 15.93%, range = 0 ~ 99.49%); 4.89% was Asian (SD = 9.55%, range = 0 ~ 82.74%); 1.41% was American Indian/Alaska Native (SD = 7.31%, range = 0 ~ 99.54); 0.36% was Hawaiian/Pacific Islander (SD = 0.95, range = 0 ~ 16.02); and 4.34% was from two or more races (SD = 4.27, range = 0 ~ 100). Also, the average proportion of students eligible for receiving free or low-priced lunch was 58.03% (SD = 25.37, range 0 ~ 100).

With relation to experience of SWPBIS or CICO, those schools had implemented SWPBIS for 5.63 years on average (SD = 3.50; range = 1 to 18), and CICO for 3.73 years on average (SD = 2.28; range = 1 to 8). The TFI T1 scores from the sample ranged from

6.67% to 100% (*M* = 81.22, *SD* = 16.36), and 81.31% met or exceeded 70% as the criterion of TFI T1. The TFI T2 scores ranged from 3.85% to 100% (*M* = 71.85, *SD* = 23.99), and 59.24% met or exceeded 70% as the criterion of TFI T2.

Measures

SWPBIS Tiered Fidelity Inventory (TFI)

The TFI includes a total of 45 items (Tier 1 = 15; Tier 2 = 13; Tier 3 = 17) across 3 tiers and 10 subscales (i.e., Teams, Intervention, Evaluation for Tier 1, 2; Teams, Resources, Support Plans, Evaluation for Tier 3). Each item is rated based on a 3-point Likert-type scale including *not implemented, partially implemented,* or *fully implemented* (Algozzine et al., 2014). Total and subscale scores each tier are calculated and divided by the maximum scores to produce the percent scores. Total, subscale and item scores of the TFI at three tiers during 2015-16 (only for Question 2) and 2016-17 school years were used depending on different research questions. Skewness were within the threshold range (e.g., TFI T1 = -1.26 ~ -1.11, Tier 1 Team = -1.04 ~ -1.07, Tier 1 Implementation = -1.21 ~ -1.10, Tier 1 Evaluation = -1.51 ~ -1.28, TFI T2 = -0.93 ~ -0.81, Tier 2 Team = -1.24 ~ -1.07, Tier 2 Intervention = -1.12 ~ -1.01, Tier 2 Evaluation = -0.55 ~ -0.48, TFI T3 = -0.58 ~ -0.46, Tier 3 Team = -0.73 ~ -0.71, Tier 3 Resources = -0.59 ~ -0.57, Tier 3 Student Support = -0.60 ~ -0.48, Tier 3 Evaluation = - 0.24 ~ -0.15), indicating that the distributions of each total or subscale indicator were weakly skewed to the left.

Student Outcome Variables

Major official discipline referrals per 100 students per day. To index overall levels of problem behaviors, major official discipline referrals (ODRs) data from SWIS for two consecutive years (2015-16 and 2016-17) were used. Specifically, major ODRs rates were standardized by dividing total events of major ODRs by total days of schools and total number of enrolled students and multiplying them by 100. Skewness of these measures were 22.18 in 2015-16, and 30.06 in 2016-17, which indicated that the distributions were seriously skewed to the right. To make these indicators closer to normal distribution, they were square rooted, generating skewness within the threshold range, 1.31 in 2015-16 and 1.85 in 2016-17 (M = 0.55 – 0.57, SD = 0.29 – 0.31).

Percent of students with 0 to 1 major ODRs. SWIS allows schools to categorize major ODR data using common cut-points (0-1, 2-5, and 6 major ODRs received), and to monitor student responses and needs to each tier. From this data, proportions of students with 0 to 1 major ODRs were calculated indicating students whose behaviors are responsive to effective implementation of SWPBIS. Skewness of these measures were -1.68 in 2014-15and -2.25 in 2015 – 2016, which indicates that the distributions were within or slightly outside the threshold. To make these measures closer to the normal distribution, an arcsine transformation was applied, producing weak skewness (range = -0.61 ~ -0.49).

Differences of referral risk index between African American and non-African American students. To measure overall racial disproportionality of major ODR rates, a risk difference was used by subtracting the referral risk index for non-African American students from referral risk index for African American students. Larger difference between two groups in their major ODRs per 100 students per day indicates more disproportionality in ODRs. Their skewness (3.22 ~ 3.74) above the threshold indicates that the distributions were skewed to the right.

Percent of students meeting or exceeding state-wide standards in reading and math tests. To index overall academic outcomes, % of students with meeting or exceeding state-wide standards in reading and math tests of two years (respectively 2014-15 and 2015-16) were included. Each state annually provides adequate yearly progress reports for schools and local educational agencies via valid and reliable standardized assessment of reading or mathematics (Kim & Sunderman, 2005). Due to variations in the instruments, state-level standards, and other relevant policies, state mean centering was applied by subtract state means from individual values. When only disaggregated data by grade levels were available, the averages of all percentages of students meeting or exceeding state standards across grades were used. All indicators showed weakly positive skewness (ranging from 0.19 ~ 0.55), which indicates that the distributions of all measures were close to a normal distribution.

Percent of Students enrolled in CICO. To quantify the proportion of students receiving CICO, % of students archived in CICO enrollment list (AND, also with CICO daily points data) during a school year (respectively 2015-16 and 2016-17) were obtained by dividing the number of students enrolled in CICO by total enrollment from SWIS. Skewness were 1.84 in 2015-16 and 3.30 in 2016-17, indicating that the distributions were skewed to the right. To make these indicators closer to normal distribution, logit transformations were made, reducing the skewness to -.60 to -.47.

Percent of students meeting 80% or more of CICO daily points. To measure overall improvements of student performance in response to CICO, the percent of students meeting or exceeding 80% daily points during a school year (respectively 2015-

16 and 2016-17) was used. Skewness were within threshold, -1.34 in 2015-16 and -1.28 in 2016-17, indicating that the distributions were weakly skewed to the left.

Major ODRs per 100 days per students enrolled in CICO. To measure overall levels of problem behaviors for students enrolled in CICO, major ODRs per 100 days per student only for students receiving CICO were used in 2015-16 and 2016-17. Skewness were 3.22 in 2015-16 and 3.16 in 2016-17, indicating that the distributions were skewed to the right. To make these indicators closer to normal distribution, square root transformations were made, reducing the skewness to 0.83 to 0.84.

Implementation and/or Contextual Variables

Administrations of the TFI during the school year. To measure how frequently TFI scores were monitored by the team, counts of entering TFI data during the school year (2016-17) were used in this study.

Total number of FOI measures administered during the school year. To measure how many fidelity tools other than the TFI are used by school teams, total number of FOI measures used during the school year were included in this study.

Counts of viewing SWIS Core or Additional Reports per year. SWIS offers the View Report dashboard to display the data so school teams can monitor school-wide patterns of problem behaviors. To measure actual data use for decision making, counts of viewing Core or Additional Reports in SWIS were used in this study. Core Reports (assisting in identification of school-wide patterns of problem behaviors; PBISApps, 2016a, September) are organized into school summary (involving all core reports), average referrals per day per month (the mean ODRs per day per month during the year), referrals by problem behaviors (the most frequently occurring behaviors), referrals by time (in what time periods of the day – 15 minute increments – referrals are most likely to occur), referrals by location (which locations in the school the referrals are the most likely to occur), referrals by day of week (what days of week referrals are the most likely to occur), referrals by grade (to what graders referrals are the most frequently issued) and referrals by student (the most frequently referred students).

To refine data to specific areas of interest, Additional Reports offer multi-year reports for average referrals per day per month, location, and problem behaviors (presenting each report across multiple years), referrals by staff (presenting staff members who referred the student; only accessible to SWIS Admin users), suspension/expulsion (informing counts and number of days of suspensions and/or expulsion events, and students suspended or expelled), school ethnicity (displaying graphs in referral risk index, referral risk ratio, students with referrals by ethnicity, and referrals by ethnicity), triangle data (the proportion of students with 0 – 1, 2 – 5, and 6 or greater referrals; unavailable in this study), and year-end data (comprehensive examinations of referral data during the entire year; unavailable in this study). Accesses to individual options under each type of reports were respectively counted and used for different models in this study.

Counts of viewing CICO-Reports during the school year. CICO-SWIS (PBISApps, 2016b, September) provides schools with the View Reports feature to assist school teams in their regular monitoring of CICO data. The View Report includes School-Wide (schoolwide data on CICO implementation per day), Average Daily Points by Student (% of awarded points during a given range of dates for individual students), Individual Student Count (% of a particular student's earned points per day during a specified range of dates), Individual Student Period (% of a particular student's earned points during each period for a specified range of dates), and Student Single Period (% of a particular student's earned points during a single period each day for a specified range of dates). Uses of individual options were counted and selected for different models.

Years of implementing SWPBIS. Years of implementing SWPBIS was measured by the proxy of the initial year of entering fidelity data in PBISApps. Except for Question 4, this continuous variable (range = 1 ~ 18) was used as a dummy variable because moderating effects of years implementing SWPBIS on fidelity effects on student outcomes were tested in this study. Although a prior study (Turri et al., 2016) categorized this continuous variable into three levels (0 – 1 years as initial implementation stage, 2 - 4 years as full operation stage, and 5 or more years as sustainability stage), there are no commonly agreed-upon division rules. Depending on the nature of student outcomes, hypothesized roles of years of implementing SWPBIS and preliminary visual exploration of data distributions, different cut-offs were applied. Thereby, years of implementing SWPBIS as a continuous variable (ranging from 1 to 18) was recoded as two dummy variables. To indicate the initial period of implementation, one variable represented whether schools had implemented SWPBIS for 4 or less years (0 =5 or more years; 1 = 1 to 4 years). Although previous studies (Turri et al., 2016; Kim et al., 2018) used the cut point at year 2 or 3, there were only small numbers of elementary and secondary schools (approximately 10.82 ~ 23.83%) in the initial period determined by this cut point in the sample (insufficient for stable convergence in analytic process), and thereby, the cut-point was extended to year 4. Looking into the schools during the initial 4 years of implementation (year 1 to 4), schools with longer years of SWPBIS implementation tended to have a stronger fidelity in TFI T1, a lower major ODRs per 100 students per day, and a larger proportion of students with 0 to 1 major ODR, whereas after 4 years of implementation, those for longer years showed a weaker fidelity in TFI T1, a higher major ODRs per 100 students per day, and a smaller proportion of students with 0 to 1 major ODR. In recognition that the sustained implementation of SWPBIS with adequate fidelity might be needed to improve academic achievement by reducing problem behaviors and creating a positive learning environment (Gage et al., 2016), another dummy was used indicating whether schools had implemented 7 or longer years (0 = year 1 to year 6; 1 = year 7 or greater) implementing SWPBIS, which was created to investigate the association between the TFI and academic outcomes (in research question 2). This cut-point (year 7) was selected by exploration of academic outcome data by schools with different years of implementation. For schools during the initial 6 years, the schools with longer experiences tended to have a higher academic achievement, whereas for those past 6 years (in year 7 or greater), tended to show lower academic achievement.

Years of implementing CICO. Years of implementing CICO was measured by the proxy of the initial year of entering student data in CICO-SWIS. Due to the sampling procedure (selecting only schools with CICO-SWIS data in 2015-16 and 2016-17), there were no schools just starting CICO implementation (year 1). Except for Question 4, this continuous variable (ranging from 2 to 8) was used as a dummy variable with two levels (0 = 4 or more years; 1 = 2 to 3 years). As the visual inspection of scatterplots was unable to capture the clear pattern in outcome data across schools with different years of CICO implementation, year 4 (= median) was used to determine the initial period of CICO implementation.

School demographic variables. School demographic variables (e.g., enrollment, ethnicity, free and/or reduced-price lunch, grade level, school type, urbanicity) as covariates were included in the analysis models to control for impacts of potential confounding variables. These variables were primarily pulled from the 2014-15 data and/or the 2015-16 from the National Center for Education Statistics (NCES). Specific demographic information was found in *Participants and Settings*.

Enrollment. Total enrollment data obtained from NCES in 2014-15 (for Question 2) or 2015-16 were mainly used as a continuous covariate to control its potential impacts on outcomes. When NCES data were not available, missing data were filled by the enrollment data from SWIS or state-level aggregated data (if accessible).

Percent of African American students. Racial enrollment data were obtained from NCES in 2014-15 and (if NCES data were not available) from SWIS and state enrollment data. Percents of African American students were calculated by dividing counts of African American students by total enrollment and converting to a percentage.

Percent of students eligible for receiving free and/or reduced-price lunch. Percent of students eligible for receiving free and/or reduced-price lunch (FRL) obtained from NCES was used as a continuous covariate. It was calculated by dividing the total counts of students eligible for FRL services by total enrollment and converting to a percentage.

Regular schools. School type variable was a 5-level categorical variable (1 = regular, 2 = special education, 3 = vocational/technical, 4 = other, 5 = reportable program) obtained from NCES. There were no other schools or reportable programs in the samples in this study. Regular school was created as a dummy variable to control its possible impacts on student outcomes.

Urbanicity. Urbanicity is a 12-level categorical variable obtained from NCES, ranging from 41 = rural/fringe, 42 = rural/distant, 43 = rural/remote to 11 = city/large, 12 = city/mid-sized, 13 = city/ small). It was recoded to a semi-continuous variable from 1 = rural/remote to 12 = city/large. Higher values indicate more strongly urbanized areas.

Grade levels. Grade level from NCES is a four-level categorical variable including primary (low grade = Prekindergarten through Grade 3, high grade = up to Grade 8), middle (low grade = Grade 04 through 07, high grade = Grade 4 through Grade 9), high (low grade = Grade 07 through 12, high grade = Grade 12), and other (https://nces.ed.gov/ccd/psadd.asp). According to the NCES definition, it is hard to clearly determine middle schools (e.g., Grade 4 – 5, Grade 7 – 9) from elementary (e.g., PreK – Grade 8) and high schools (e.g., Grade 7 – 12). Therefore, two binary categorical variables (1 = elementary, 2 = secondary) was created by combining middle and high schools into secondary school category and moving Grade 6 to 12 schools from other to secondary school category and separately creating a middle school (indicating inclusion of grade 6 – 8) dummy variable. The binary categorical (elementary vs. secondary) variable was used to model the multi-group analyses, and the middle school variable was included as a covariate each group.

Procedure

Data Collection Procedure

Two datasets obtained from an IRB-approved data repository at the University of Oregon were used depending on types of student outcomes. First, the preliminary sample included 1,673 schools from 30 states in United States that completed the TFI and school discipline records (from SWIS) or CICO intervention data (from CICO-SWIS) in the 2015-2016 school year via PBISApps. However, only 284 schools had school discipline outcomes of the 2014-15 school year, the important covariates controlled for. Out of 603 schools with CICO-SWIS data, only 154 schools had the prior year outcomes. To increase the sample size, a sequential cohort design was used by adding another dataset involving 2,380 schools with the TFI and SWIS and/or CICO-SWIS data in the 2016-17 school year for research questions addressing the association between the TFI and behavioral outcomes.

In the analytic software used in this study (*M*plus), cases missing independent variables are removed via list-wise deletion, so only schools with the prior-year outcome data were selected. Specifically, 2,338 schools had TFI T1 scores and SWIS outcome data in 2016-17, in which 1,716 schools had SWIS outcome data in 2015-16 as well. In addition, the 24 schools including both elementary and secondary grade levels (e.g., pre-kindergarten to 12th grade) were excluded for the group-based models. For Question 1, this study focused on 1,692 schools. With relation to Question 3, 1,774 schools (74.5%) that reported TFI T2 scores in 2016-17 were initially filtered out, in which 841 schools had CICO-SWIS data in 2016-17. A total of 573 schools with the prior year outcomes were selected including three schools with both elementary and secondary graders because insufficient sample size of secondary schools made it difficult to use a multi-group approach. For Question 4 regarding counts of reviewing SWIS and/or CICO-SWIS reports, 2,379 schools with TFI T1 scores and SWIS data or with TFI T2 scores and CICO data in 2016-17 were used.

Academic achievement data were collected from the state-level database readily accessible across states in 2014-15 and/or in 2015-16 for the initial sample of 1,673 schools because of limited data availability in 2016-17. State data centers across the 30 states with 5 or more schools included in this dataset were visited to collect percentage of students meeting or exceeding state-level standards in reading or math achievement tests. First, school-level aggregated data for all students from all grades were selected in priority. If only disaggregated data across grade levels were available, the average across grades were calculated to obtain the school outcome measure. If school-level data are not accessible, school report cards were individually drawn. Otherwise, the data were requested directly from the states. Data from several states were not included in analyses due to the changeover to a new assessment measure (e.g., Arizona, New York), a dual assessment system that allowed schools to choose one of assessment measures (e.g., Massachusetts), or largely missing data in reading and math in 2014-15 for unknown reasons (e.g., Connecticut, Pennsylvania). Out of 1,386 elementary or secondary schools, from 16 states, with both the TFI and academic data in 2015-16, 1,361 schools reported academic data (either in reading or math) in the prior year, where were selected for the second research question regarding academic outcomes. A double check for accuracy of data entry was made with 50% of data, finding that 99.7% were correct.

Analytic Procedures

Structural equation modeling was applied using M*plus*7 (Muthén & Muthén, 1998-2012). To account for differences in schools by grade levels, two parallel models were built on across two groups, elementary and secondary schools, via the "GROUPING" option in "VARIABLE" command. However, depending on the sample sizes and model complexity across different research questions (e.g., Question 3, 4), grade levels served was included as a covariate.

Initial measurement models. Two measurement models for the latent fidelity variables were tested for their invariance by gradually constraining equivalence of the parameters (factor loadings, manifest variable intercepts and residual variance) from all freely varying parameters to all equally constrained parameters between two groups (Byrne, 2013). To determine degree of group invariance of the measurement models from the study datasets, model fits and composite reliability were compared between configural invariance (as a baseline model) and other stricter invariance models such as metric or scalar invariance (Chen, Sousa, & West, 2005; Vandenberg, & Lance, 2000). For instance, configural invariance was specified by modeling the same number of factors and relevant indicators without any equality constraints between the two groups. Metric invariance model with equality constraints on factor loadings were tested to determine if the construct measured by the TFI does have same meaning across two groups. If the metric invariance model had better model fits than the configural, the scalar invariance model where factor loadings, and intercepts of the observed indicators were equal between groups was additionally modeled to answer whether the meaningful comparisons of factor means would be possible by establishing the common origin and unit of measurement (Chen et al., 2005). When the scalar invariance model showed more suitable than the metric, residual invariances of indicators under each factor were additionally constrained to be same across group. Establishing this strictest invariance model indicates that score differences between elementary and secondary schools in the

TFI are explained by only differences between grade levels on the latent factors (Burns, Walsh, Gomez, & Hafetz, 2006).

Inspection of multi-collinearity. With the optimal measurement models, preliminary structural models involving the TFI and each of the outcome variables were specified to inspect the multi-collinearity in regression. First, the three factors, TFI Tier 1, 2, and/or 3 scores, were specified as primary predictors. On these co-varying three factors, individual dependent variables were regressed one by one. Because it is not available to obtain the variance inflation factor (VIF) in M*plus*, alternative models with the only one factor corresponding to the research question (e.g., TFI Tier 1 factor for Question 1) were specified and compared with different models with the single factor of the main interest (e.g., tier 1 in Research Question 2) and student outcomes. For example, if the effect of one factor were significant in the single factor model but not in the three-factor model, multi-collinearity between factors would be highly suspected. If the multi-collinearity was considered serious enough to affect the parameter estimation, the one factor model involving the most relevant subscale or item scores was selected as a final measurement model. Extended from the final measurement model, the hypothesized model involving all predictors and covariates were specified.

Specification and estimation of main and interaction effect models. The effects of demographic covariates on outcome variables were controlled involving prior year outcomes in correspondence to each of dependent variables, and relevant contextual variables across all models. Step-wise regression was applied for covariates in the models, in which only statistically significant covariates were retained. Under the multi-group models, structural parameters (e.g., regression path coefficients, factor means) were cumulatively constrained equal if grade-levels (elementary vs secondary) was not attributed to some degrees of variance in path coefficients (Byrne, 2013). A Wald test was used to determine the necessity of equality constraints, and significance of moderating effects of grade-levels. Except the hypothesized predictors, moderators, and dependent variables of the interests, only significant parameters (associated covariates) in either primary or secondary were included in both models.

Two structural models with and without interaction terms (fidelity latent scores by years of implementation) were made for the following practical purposes: a) to guide the modeling of interaction models, and b) to determine the moderation effect of grade levels under the main effect models without heavy computation process for interaction terms. Specifically, years of implementing SWPBIS as a dummy variable was included not only as a sole covariate, but also as an interaction term between years of implementation and the fidelity factors. To define the interaction term between the fidelity factors as continuous latent variable and years implementing SWPBIS as a dichotomous variable, the XWITH command was used under TYPE = RANDOM and ALGORITHM = INTEGRATION.

Also, in a recognition of the feature of the TFI in its usage, the sensitivity test was conducted by comparing the results from the schools with at least one score at the tier of the main interest to those from the schools with all TFI scores of three tiers (i.e., list-wise deletion of schools). In addition, from these schools with all scores of three tiers, total TFI factor score were estimated using three percentage scores of three tiered scales.

Nested data structure. In all datasets, schools were nested within districts, and schools and districts were nested within states. Intra-class correlations were calculated for all outcome measures under the unconditional models indicating significant clustering effects at district (e.g., ICC = .349 for major ODRs per 100 students per day; .313 for % of students with 0 – 1 major ODR; .102 for difference of racial risk index; .441 for state mean-centered reading; .422 for state mean-centered math; .215 for logit-transformed percent of students with CICO daily point data; 0.231 for percent of enrolled students with 80% or greater goal attainment; .240 for square-root of major ODRs per 100 days per day) compared to those at state levels (e.g., ICC = .061 for major ODRs per 100 students per day; .045 for % of students with 0 – 1 major ODR; .011 for difference of racial risk index; .001 for reading; .001 for math; .160 for logit-transformed percent of students with CICO daily point data; ; .033 for percent of enrolled students with 80% or greater goal attainment; .100 for square-root of major ODRs per 100 days per day). Therefore, all prediction models filtered out only the clustering effects of district via "TYPE = COMPLEX" which makes an adjustment to standard errors to account for unequal subject selection probabilities and non-independence throughout M*plus*7 (Muthén & Muthén, 1998-2012; Muthen & Satorra, 1995; Roland & Thomas, 2015). However, 10% or more of variations in those outcome measures associate with CICO data were explained at the state level, and thereby, state variables were included as covariates if needed. This solution was chosen instead of straightforward specification of multi-level models (despite similar functions) for parsimony in a sense that district-level estimation is not of main interest in this study.

Model (re)estimation and modification. The maximum likelihood estimation with robust standard errors and Chi-square tests (ESTIMATOR=MLR) that are robust to non-normality and non-independence of observation was used to compute standard errors via a sandwich estimator (Muthén & Muthén, 1998-2012). If item indicators were used for the latent factors, the robust weighted least squares (ESTIMATOR=WLSMV) was used to accurately estimate three-categorized ordinal data, which is not normally distributed as well (Brown, 2006). For each model, multiple model fit statistics were used with the following criteria for determining whether a particular analytic model is acceptable for the dataset (Hooper, Coughlan, & Mullen, 2008; Yu, 2002): a) χ^2 low enough for $p > .05$ based on degrees of freedom, b) Comparative fit index (CFA) > .95, c) Tucker-Lewis Index (TLI) > .95, d) standardized root mean square residual (SRMR) < .08, and e) Root mean square error of approximation (RMSEA) and its 90% Confidence Interval < .08 as well as f) Weighted Root-mean-square Residual (WRMR) < 1. However, if the interaction terms are introduced to models, these model fit indices are not available without sufficient means, variances, and covariances for model estimation because there are no mean, variance or covariances between the latent interaction terms, and other parameters, and under the matrix (Maslowsky, Jager, & Hemken, 2015). Instead, a log-likelihood ratio test was obtained from and compared for both models (Maslowsky et al., 2015). During the modeling procedure, modification indices were used to improve the model fit of model to acceptable level, and a few co-variances of error term within each factor were added only when it is theoretically or contextually explainable. Each modification was comparatively examined one by one.

Missing data. If a school does not complete an entire tier of the TFI (as is sometimes recommended for multiple administrations during a school year), the TFI data with a score of zero for all items in a specific tier were coded as missing. In this study, only schools with TFI scores of the relevant tiers and outcome data across two years were selected, which generated no missing in critical study variables. Although merging multiple datasets produced missing, there were only a few (less than 5%). Thereby, missing data were addressed via full-information maximum likelihood estimation (using TYPE = MISSING H1 as a default) with the WLSMV or MLR estimator because it is useful when there are no or small amount of missing data or when missing at random assumption with auxiliary covariates controlled for are plausible (Asparouhov & Muthén, 2010; Schafer, 1999).

Hypothesized Structural Models

Research Question 1. Figure 1 displays the hypothesized associations between TFI Tier 1 scores (as independent variables), and school-level student behavior outcomes (as dependent variables). Dependent variables include 1) square root of major ODRs per 100 students per day (major ODR rates), 2) difference of referral risk index of between African American students and non-African American students, 3) % of students with 0 – 1 major ODRs. For instance, major ODR rates were assumed to negatively associate with TFI T1 scores when other covariates such as school demographics and years of implementation (as a sole variable) were held equal across schools. In particular, the prior outcome of the 2015-16 year were regressed on the outcome of the 2016-17 to control the autoregressive effect. Also, the interaction term between the fidelity factor and years of implementation was included to evaluate whether the fidelity effects would vary depending on years of implementing SWPBIS.

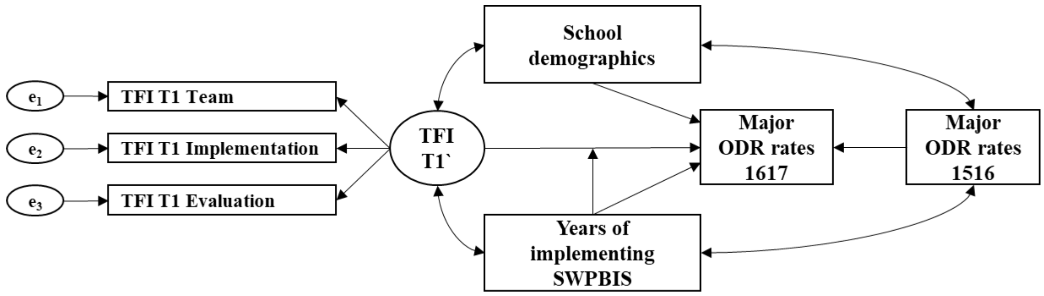

Figure 1. Hypothesized associations between TFI scores and ODRs per 100 students per day. These hypotheses were consistently applied to associations between TFI scores and other behavioral outcomes.

Research Question 2. Research question 2 investigated associations between TFI Tier 1 scores (as independent variables), and school-level student academic outcomes (as dependent variables). Dependent variables were proportions of students meeting or exceeding the state-level standards in reading and math test scores. Figure 2 illustrates the hypothesized models on the associations between TFI T1 scores and reading outcomes. These models were built for both elementary and secondary schools.

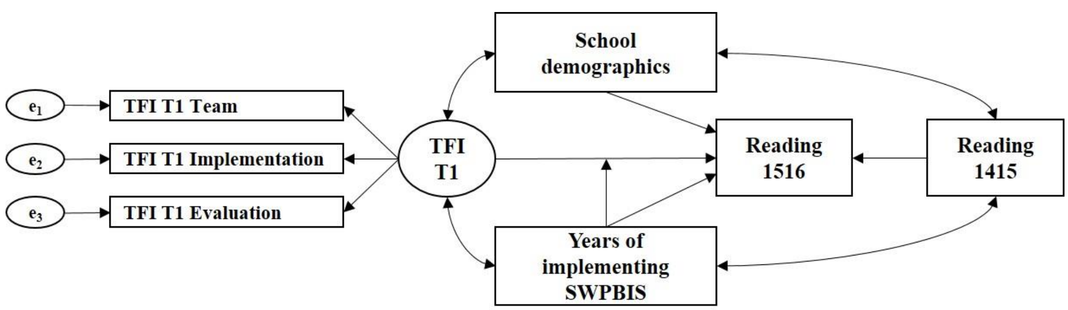

Figure 2. Hypothesized associations between TFI scores and proportions of students meeting or exceeding state-level standards in reading achievement tests. These hypotheses were consistently applied to associations between TFI scores and math achievement data.

Research Question 3. Research question 3 investigated associations between TFI Tier 2 (as independent variables), and behavior outcomes for students enrolled in

CICO (as dependent variables). As indicated by Figure 3, TFI Tier 2 scores were hypothesized to predict 1) proportion of students enrolled in CICO, 2) students meeting 80% or more daily points, and 3) major ODRs per 100 students per day for students receiving CICO in the school year of 2016-17, after controlling for corresponding outcomes in 2015-16 and other contextual variables. Fidelity effects were allowed to vary depending on years of implementation of CICO. Each of the behavior outcomes was examined in one model at a time.

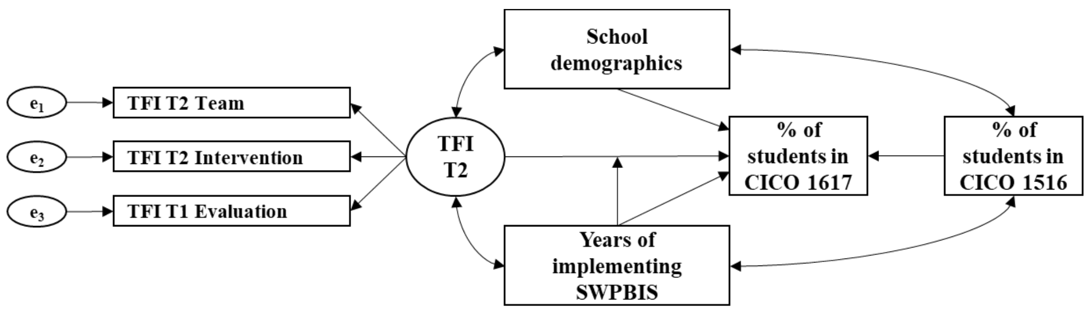

Figure 3. Hypothesized model on association between TFI scores and proportions of students enrolled in CICO. These hypothesized associations between outcome and other study variables were consistently applied to different behavioral outcomes (e.g., ODRs per 100 students per day for students in CICO).

Research Question 4. Research question 4 evaluated the concurrent validity of TFI scores with other relevant measures, such as data use activities (measured by counts of viewing SWIS Core or Additional Reports, and counts of viewing CICO Reports), ongoing assessment of fidelity (measured by counts of TFI administrations, and number of fidelity measures used during 2014-15), and years of SWPBIS implementation. As indicated by Figures 4 and 5, the TFI Tier 1 Evaluation subscale and TFI Tier 2 Evaluation subscale factor scores were defined via relevant item indicators, and they were hypothesized to covary with counts of TFI administrations, number of fidelity measures, counts of viewing SWIS Core Reports, and counts of viewing SWIS Additional Reports or counts of viewing CICO Reports. First, SWIS Core Reports were measured by eight separate count variables, including school summary, average referrals per day per month, referrals by time, referrals by location, referrals by day of week, referrals by problem behaviors, referrals by grade, and referrals by student. Second, SWIS Additional Reports include multi-year reports of average referrals per day per month, multi-year reports of locations, multi-year reports of problem behaviors, referrals by staff, school ethnicity, and suspensions. Third, CICO Reports include School-Wide, individual student count, average daily points by student, individual student period, and individual student single period. These count variables were included as observed variables.

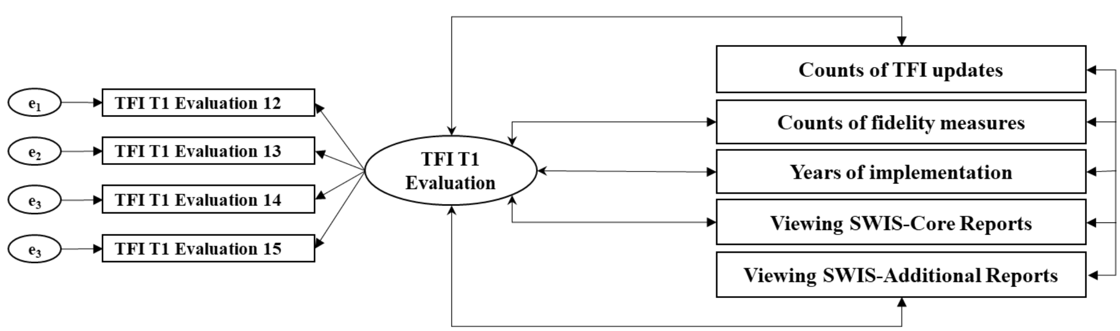

Figure 4. Hypothesized associations between TFI evaluation subscale scores of Tier 1 and relevant observation data on data use or other evaluation activities.

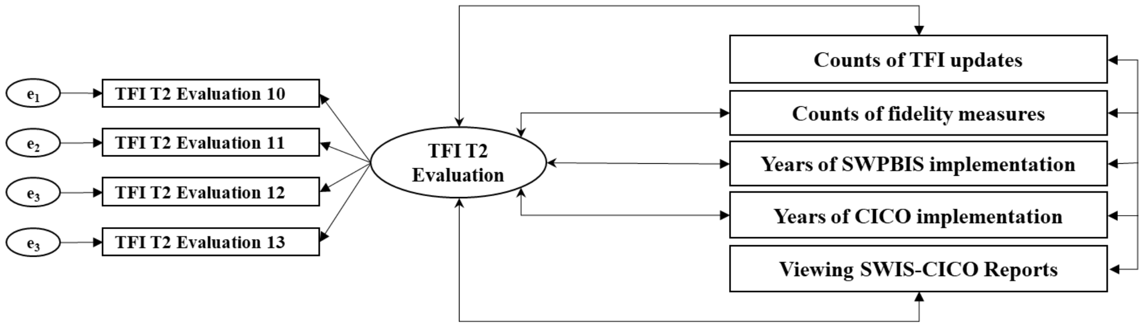

Figure 5. Hypothesized associations between TFI evaluation subscale scores of Tier 2 and relevant observation data on data use or other evaluation activities.

CHAPTER III

RESULTS

Research Question 1

Preliminary analyses were initially conducted to determine the optimal solutions to address the violations to analytic assumptions, and to revise the hypothesized measurement and structural models. Then, the results of descriptive statistics and final structural equation models were reported.

Preliminary Analyses. A three-factor measurement model, indicating Tier 1, 2, and 3 TFI scales, was initially specified using Confirmatory Factor Analysis (CFA), in which observed indicators relate to the underlying latent variables (or factors; Kline, 2005). Specifically, three factors were specified with subscale scores, respectively 3, 3, and 4 subscales for each of the Tier 1, 2, and 3 TFI factors because of the model complexity. Then, the invariance of factor structures was tested across two groups via judging the model fits to the dataset used for each of different models separately based on the pre-determined criteria. Under two groups, four different measurement models were compared with increasing degrees of invariance.

Table 3 indicated that the metric invariance model showed the best fit in terms of Chi-square tests. However, with relation to other fit indices, all four invariance models showed suitable fits to the data, meeting the pre-determined criteria. For parsimonious models, the strictest invariance model was initially selected, where factor loadings, intercepts and residual variances of observed indicator were constrained to be equal between elementary and secondary schools.

Table 3

Model Fit Statistics for Multi-Group Models of Measurement Invariance

Invariance	Chi-square		CFI	TLI	SRMR	RMSEA	
	Df	Value				Value	90% CI
Configural	64	211.65*	.972	.960	.035	.052	.045 – .060
Metric	71	218.40*	.972	.964	.039	.050	.042 - .057
Scalar	78	245.85*	.968	.963	.043	.050	.043 - .058
Residual	88	279.11*	.963	.962	.053	.051	.044 - .057

Note. $^{*}p < .05$., $^{**}p < .01$.

Table 4 presents the parameter coefficients estimated from the strictest invariance model. All factor loadings were statistically significant, ranging from .593 to .913. Internal consistency for three factors (T1 = .78 ~ .83, T2 = .87 ~ .90, T3 = .91 ~ .92) were acceptable (greater than .70). Under the most restrict equality constraints, factor means were fixed at zero in elementary schools as the reference group whereas factor means in secondary schools were freely estimated to test the mean differences between groups. As indicated by Table 4, secondary schools were shown to have significantly lower mean factor scores in TFI Tier 1, Tier 2, and Tier 3.

***Inspection of multicollinearity*.** As early reported, the three TFI scale factors (Tiers 1, 2, and 3) moderately to strongly covaried, which generated suspicions of multi-collinearity under the regression path models. To confirm whether multi-collinearity was serious enough to affect the parameter estimation, the preliminary structural models were specified with academic in 2015-2016 and behavioral outcomes in 2016-2017 measures and the prior outcomes as shown in Table 5.

Table 4

Results of Multi-Group Three Factor Models of Measurement Invariance (*N* =1,691)

Parameter			Elementary			Secondary		
			N	Standardized coefficient	*SE*	*N*	Standardized coefficient	*SE*
T1	Teams	Factor loading	1218	0.593**	0.024	473	0.652**	0.029
		Intercept		4.027**	0.125		3.393**	0.130
		Residual		0.648**	0.028		0.575**	0.037
	Implementation	Factor loading	1218	0.844**	0.019	473	0.878**	0.016
		Intercept		4.966**	0.188		4.428**	0.180
		Residual		0.288**	0.032		0.229**	0.029
	Evaluation	Factor loading	1218	0.786**	0.020	473	0.829**	0.019
		Intercept		5.003**	0.206		4.522**	0.199
		Residual		0.382**	0.031		0.312**	0.031
	Mean			0	0		-0.288**	0.054
	Composite Reliability			0.789			0.833	
T2	Teams	Factor loading	964	0.772**	0.019	345	0.809**	0.020
		Intercept		3.649**	0.138		3.375**	0.140
		Residual		0.404**	0.030		0.346**	0.032
	Intervention	Factor loading	964	0.892**	0.010	345	0.913**	0.012
		Intercept		3.339**	0.157		3.017**	0.138
		Residual		0.204**	0.018		0.166**	0.022
	Evaluation	Factor loading	964	0.837**	0.014	345	0.866**	0.013
		Intercept		2.234**	0.096		2.041**	0.083
		Residual		0.299**	0.024		0.250**	0.023
	Mean			0	0		-0.330**	0.058
	Composite Reliability			0.873			0.898	

Table 4 continued

Parameter			Elementary			Secondary		
			N	Standardized coefficient	*SE*	*N*	Standardized coefficient	*SE*
T3	Teams	Factor loading	591	0.842**	0.017	224	0.859**	0.016
		Intercept		2.453**	0.113		2.326**	0.102
		Residual		0.291**	0.028		0.261**	0.028
	Resources	Factor loading	591	0.842**	0.017	224	0.859**	0.015
		Intercept		2.257**	0.107		2.140**	0.093
		Residual		0.291**	0.029		0.262**	0.026
	Student Support	Factor loading	591	0.826**	0.018	224	0.845**	0.017
		Intercept		2.025**	0.093		1.924**	0.082
		Residual		0.317**	0.030		0.286**	0.029
	Evaluation	Factor loading	591	0.849**	0.016	224	0.866**	0.014
		Intercept		1.729**	0.079		1.638**	0.068
		Residual		0.279**	0.026		0.250**	0.023
	Mean			0	0		-0.297**	0.078
	Composite Reliability			0.905			0.917	
Factor Covariances		T1 ↔ T2		0.664**	0.045		0.644**	0.060
		T2 ↔ T3		0.606**	0.039		0.602**	0.055
		T1 ↔ T3		0.401**	0.062		0.333**	0.070

Note. The measurement models between elementary and secondary schools were constrained to be equal in their (unstandardized) factor loadings, and intercepts and residuals of observed subscale indicators. Also, the factor variances were fixed at 1 to identify the model.

$^{*}p < .05.$, $^{**}p < .01.$

Table 5.

Comparison of standardized path coefficients between 3-factor and 1-factor model in behavioral outcomes for schools implementing SWPBIS

		Standardized coefficients			
		3-factor model		1-factor model(T1)	
Outcome	Predictors	Elementary	Secondary	Elementary	Secondary
Square root of major ODRs per 100 students a day (ODR rates) in 2016-17	T1	-0.020	0.035	-0.038*	0.029
	T2	-0.062	0.016		
	T3	0.060	-0.047		
	Previous outcome	0.847**	0.904**	0.847**	0.904**
Arcsine of % of students with 0 – 1 major ODR (0 – 1 major ODR) in 2016-17	T1	0.014	-0.031	0.022	-0.033
	T2	0.037	-0.036		
	T3	-0.041	0.067		
	Previous outcome	0.846**	0.821**	0.846**	0.821**
Racial gap in ODR rates between African American and the rest of students	T1	-0.063	0.067	-0.075*	0.037
	T2	-0.101	-0.030		
	T3	0.143	-0.022		
	Previous outcome	0.556**	0.795**	0.560**	0.795

Note. $^{*}p < .05.$, $^{**}p < .01.$

As shown in Table 5, there were no significant associations between factorial scores and either outcome in 2015-2016 across two levels of schools under the three factor models. However, one factor models only with TFI Tier 1 showed significant effects of TFI Tier 1 on square root of major ODRs per 100 students per day (major ODR rates) in elementary schools, consistent with the result of the one factor model with TFI Tier 2 only (B = -.040, p = .040 in elementary schools). Similarly, TFI Tier 1 (as the sole IV) was significantly associated with difference between African American and the rest of students in major ODR rates in elementary schools (not in TFI T2 or T3 only models). Overall, the shared variance among factors in outcome measures were detected, and therefore, TFI Tier 1 factor alone was selected for the final models.

***Measurement models*.** A 1-factor model was specified with varying equality constraints and compared as shown in Table 6. The most parsimonious model for the 1-factor model showed a reasonable model fit. However, before finalizing the measurement model, the preliminary structural models were specified with interaction terms requiring heavy numerical integrations, which faced frequent failures to converge with relation to the latent fidelity score. In a recognition of the possibility of misspecification, equality constraints were gradually released in a stepwise manner. Then, the partial invariance models with equality constraints on all factor loadings and intercepts and partially on residual variances of evaluation were determined as a final measurement model of TFI T1. The result of final model (factor loadings = .61 ~ .85, composite reliability = .80 ~ .81) is reported in Table 7. The final model was run again with the selected sample of 584 elementary and 222 secondary schools with all TFI T1, T2, and T3 scores, showing an excellent fit as well: $\chi^2(5) = 4.674$, $p = .457$, CFI = 1.00, TLI = 1.00, SRMR = .047, RMSEA = .00 (CI = .00 ~ .07).

Table 6

Model Fit Statistics for Multi-Group Models of Measurement Invariance

Invariance	Chi-square		CFI	TLI	SRMR	RMSEA	
	df	Value				Value	90% CI
Metric	2	3.183	.999	.998	.033	.026	.000 - .078
Scalar	4	15.875*	.992	.988	.045	.059	.031 - .091
Residual	7	28.972*	.985	.987	.081	.061	.039 - .085
Releasing residuals							
- teams	6	23.937*	.988	.988	.055	.059	.036 - .085
- implementation	6	19.634*	.991	.991	.074	.052	.027 - .078
- evaluation	6	27.265*	.986	.986	.095	.065	.041 - .090
- teams & implementation	5	15.295*	.993	.992	.038	.049	.022 - .079
- teams & evaluation	5	22.587*	.988	.986	.072	.064	.039 - .093
- implementation & evaluation	5	20.161*	.990	.988	.078	.060	.034 - .088

Note. Configural invariance model was just identified and cannot be evaluated via Chi-square test. Metric invariance model fixed the factor means at zero in both groups whereas scalar and residual invariance models fixed the factor mean only in secondary schools.

$^{*}p < .05.$, $^{**}p < .01.$

Table 7

The Result of Revised Measurement Model

Parameter			Elementary			Secondary		
			N	Standardized coefficient	*SE*	*N*	Standardized coefficient	*SE*
T1	Teams	Factor loading	1218	0.607**	0.024	473	0.617**	0.031
		Intercept		4.100**	0.129		3.655**	0.159
		Residual		0.631**	0.029		0.620**	0.038
	Implementation	Factor loading	1218	0.838**	0.022	473	0.817**	0.025
		Intercept		5.025**	0.199		4.302**	0.178
		Residual		0.298**	0.038		0.332**	0.041
	Evaluation	Factor loading	1218	0.816**	0.020	473	0.849**	0.020
		Intercept		4.980**	0.207		4.550**	0.207
		Residual		0.334**	0.033		0.279**	0.034
	Mean			0	0		-0.274**	0.055
	Composite reliability			0.802			0.810	

Descriptive Statistics. Table 8 indicates the means, and standard deviations of study variables for elementary and secondary schools in Research Question 1. Overall, TFI subscale scores tended to decrease from Tier 1 to Tier 3. Regarding school grade levels, elementary schools showed higher mean fidelity scores than secondary schools across tiers and subscales. With relation to outcome measures, secondary schools showed higher levels of problem behaviors and racial disproportionality than elementary schools.

Table 8

Descriptive Statistics of Study Variables

	Elementary			Secondary		
	N	*M or %*	*SD*	*N*	*M or %*	*SD*
TFI Tier 1 Team (%)	1218	83.64	20.11	473	79.33	23.58
TFI Tier 1 Implementation (%)	1218	83.33	16.51	473	77.44	19.02
TFI Tier 1 Evaluation (%)	1218	85.33	17.33	473	82.45	18.84
TFI Tier 1 (%)	1218	83.91	15.10	473	79.03	17.32
TFI Tier 2 (%)	964	76.35	21.66	345	69.06	24.14
TFI Tier 3 (%)	591	65.48	26.01	224	58.22	28.02
Square root of major ODRs per 100 students per day in 1516	1204	0.54	0.28	462	0.65	0.53
Square root of major ODRs per 100 students per day in 1617	1204	0.55	0.30	462	0.67	0.55
Proportions of students with 0 – 1 major ODR in 1516	1217	91.07	7.23	462	87.57	10.05
Proportions of students with 0 – 1 major ODR in 1617	1217	90.74	7.48	462	87.01	10.59
Racial gap of referrals per student per 100 days between African American and the other students in 1516	635	0.49	0.74	265	0.68	0.83

Table 8 continued

	Elementary			Secondary		
	N	*M or %*	*SD*	*N*	*M or %*	*SD*
Racial gap of referrals per student per 100 days between African American and the other students in 1617	635	0.48	0.66	265	0.63	0.78
% of schools implementing for 4 or fewer years	1218	30.87	46.22	473	35.94	48.03
Mean degrees of urbanicity (1=the most urban ~ 12 = the most rural)	1177	7.54	3.38	451	6.94	3.33
% of students eligible for free or reduced lunch	1175	58.06	25.53	451	52.14	22.04
% of African American students	1210	10.39	16.99	471	8.88	14.61

The cut-off point of Year 4 was selected on the years of implementing SWPBIS based on the visual inspection of fidelity scores and student outcomes across schools with varying years of implementing SWPBIS as detailed in the section below.

TFI T1 scores across years of implementing SWPBIS. Similar to previous research (Nese et al., 2018), schools implementing SWPBIS for two or more years (in Year 2 or more) tended to exceed 80% or above of TFI T1 fidelity as shown in Figure 6. Then, schools in Year 5 to 8 showed relatively higher scores more or less than 83%, and afterword (Year 10 to 12), schools with increasing years of implementing SWPBIS tended to have lower (but still above the criterion, 70%) fidelity scores.

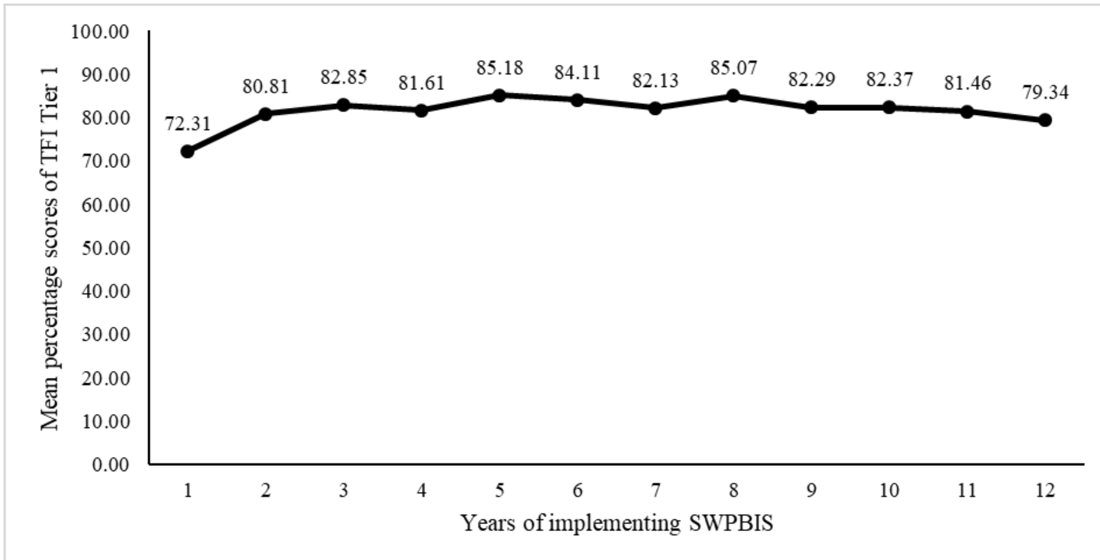

Figure 6. Mean-percentage scores in TFI Tier 1 over years implementing SWPBIS in 2016 – 2017. Schools implementing for 13 or more years were not included due to small sample sizes (n = 1 ~ 23).

Behavioral outcomes across years of implementing SWPBIS. Figure 7 indicates that elementary schools during the initial 4 years of implementation showed a pattern of fewer in problem behaviors, and those during 5 or longer years showed higher over their experiences. In a similar manner, Figure 8 indicates that schools during the initial 4 years (1 – 4 years) display higher in the arcsine of proportions of students with 0 – 1 major ODR and those past the initial period reported slightly lower over years. Along with the immediately growing and slowly declining fidelity, Figure 7 and 8 offered the hypotheses that 1) the implementation of SWPBIS with stronger fidelity would improve the school-wide behavioral outcomes from schools during the initial period (year 1 ~ 4); and 2) the association of fidelity and student outcomes would gradually be lower in strength afterwards. To observe such a hypothesized pattern from the data, scatterplots with outcomes and TFI T1 percentage scores for three types of schools in varying years of implementation (year 1 ~ 4, year 5 ~ 8, year 9 or longer) were visually compared. Overall negative (or slightly positive) associations between major ODR rates and TFI T1 scores, and positive (or slightly negative) associations between proportions of students with 0 – 1 major ODRs and TFI T1 scores were found from all types but with the strongest extent from schools during the initial period. However, disproportionality measures were not considered with relation to years of implementation because of relatively small size of subsample and ongoing fluctuations (inconsistency).

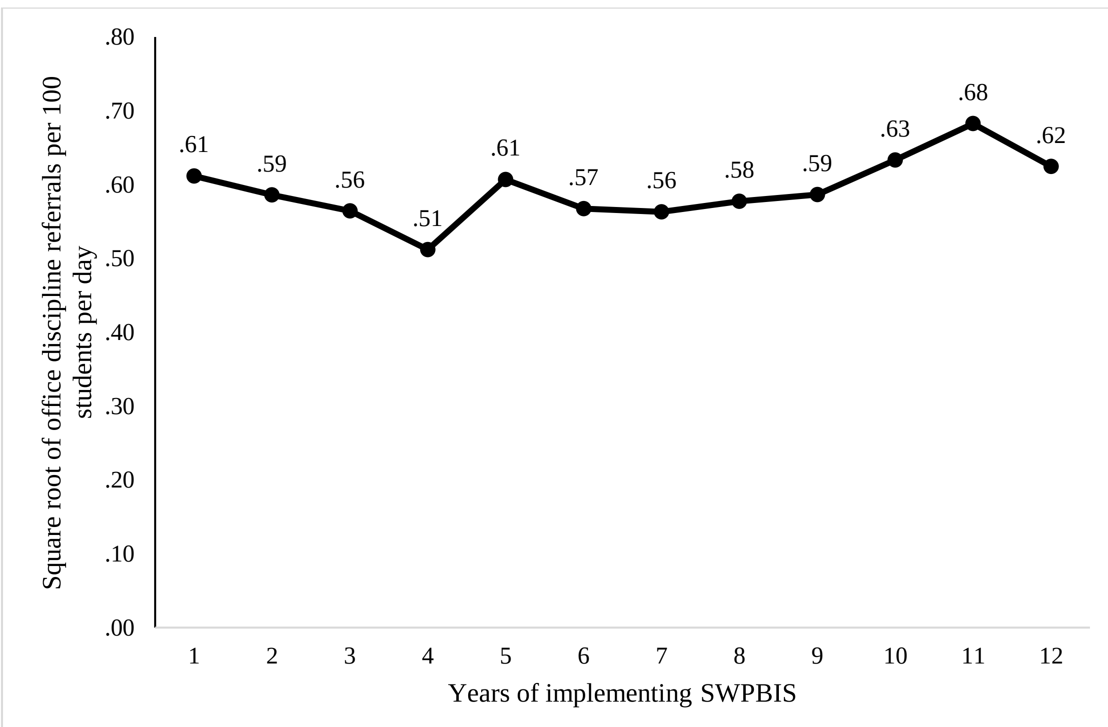

Figure 7. Square root of major root of major office discipline referrals per 100 students per day over years. Schools implementing for 12 or more years were not included due to small sample sizes (*n* = 1 ~ 23).

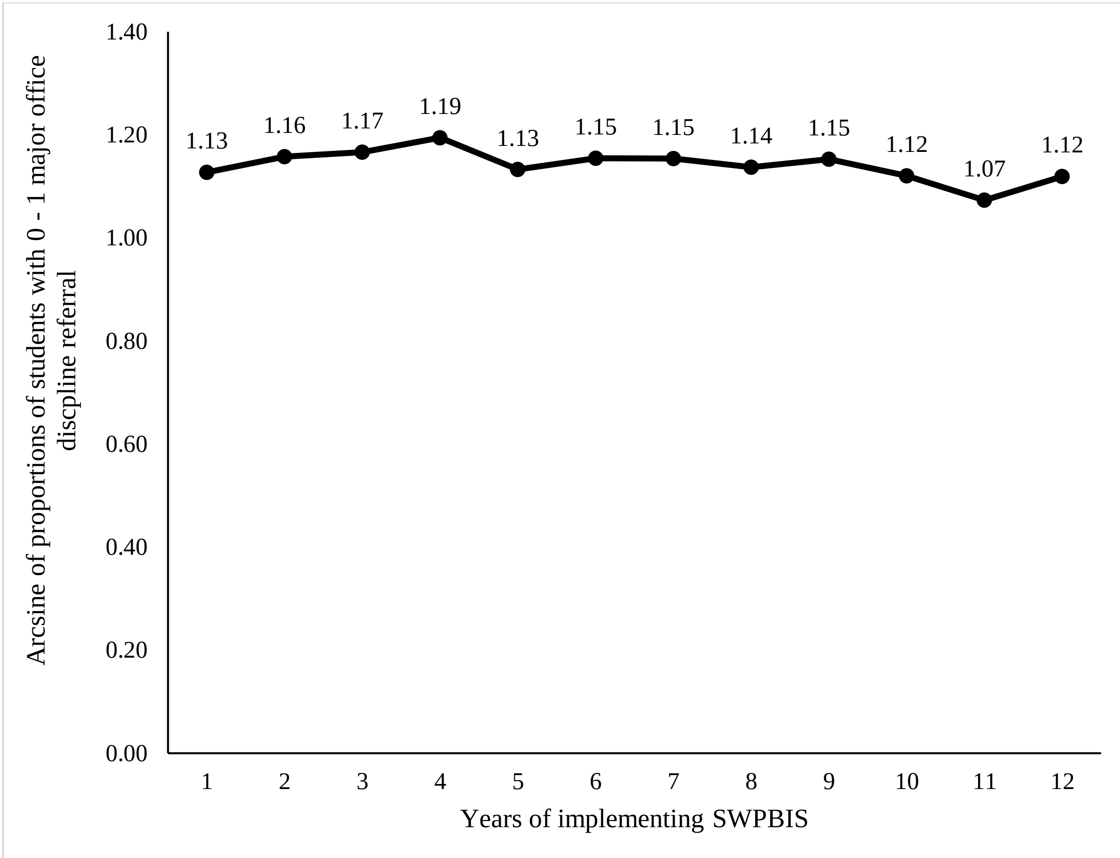

Figure 8. Arcsine of the proportion of students with 0 – 1 major office discipline referral. Schools with 12 or longer years were not involved due to small sample sizes.

Results of Structural Models. Table 9 compared the model fits of between interaction and main effect models. For all outcome measures, the loglikelihood ratio tests were conducted, finding adding interaction terms did not result in statistically significant differences in goodness of fit. As there were no statistically significant interaction terms as indicated by Table 10, the null hypotheses of no moderating effects of years implementing on fidelity effects were not rejected. Thereby, the sensitivity tests were followed based on the assumption that the fidelity effects would not significantly vary even past four years of SWPBIS implementation.

Table 9

Model Fit Statistics of Final Structural Models for School Wide Behavioral Outcomes

Model fit indices	Square root of major ODRs per 100 per day		Proportions (%) of students with 0 – 1 major ODRs		Racial gap in major ODRs per 100 per day	
	Main effect Model	Interaction effect model	Main effect model	Interaction effect model	Main effect model	Interaction effect model
Loglikelihood						
Free parameter	21	23	22	24	22	24
H0 Scaling correction factor	3.132	2.956	2.082	2.015	2.8753	2.7446
H0 Value	-19565.738	-19565.414	-24959.480	-24959.030	-11900.355	-11899.227
CFI	.996		.983		.961	
TLI	.996		.980		.953	
SRMR	.052		.054		.060	
RMSEA value	.039		.039		.065	
RMSEA 90% CI upper bound	.052		.050		.081	
AIC	39173.476	39176.828	49962.960	49966.061	23844.711	23847.027
BIC	39286.462	39300.574	50081.530	50095.410	23950.364	23967.087

Note. Mean scores of the latent fidelity scores were constrained at zero in both groups.

Table 10

Results of Final Structural Models for Behavioral Outcomes

Parameter		Schools with TFI Tier 1 scores			
		Elementary		Secondary	
		Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Model 0: Main effect models					
Square root of major ODRs per 100 students per day	Prior year outcome (e)	0.945^{**}	0.052	0.945^{**}	0.052
	TFI T1	-0.001	0.000	0.001	0.001
	Year 1 to 4 in implementing SWPBIS (e)	0.028^{*}	0.011	0.028^{*}	0.011
	Urbanicity (e) - 1 = the most urban	-0.004^{*}	0.002	-0.004^{*}	0.002
Proportion of students with 0 – 1 major ODR	Prior year outcome (e)	0.850^{**}	0.026	0.850^{**}	0.026
	TFI T1	0.016	0.011	-0.031	0.021
	Year 1 to 4 in implementing SWPBIS (e)	-0.382	0.262	-0.382	0.262
	% of students eligible for free-/reduced price lunch(e)	-0.011^{*}	0.004	-0.011^{*}	0.004
	Urbanicity (e) - 1 = the most urban	0.109^{**}	0.038	0.109^{**}	0.038
Difference in major ODR rates of between African and non-African students	Prior year outcome (e)	0.578^{**}	0.104	0.578^{**}	0.104
	TFI T1	-0.004^{*}	0.002	0.001	0.003
	Year 1 to 4 in implementing SWPBIS (e)	0.011	0.035	0.011	0.035
	Middle school (grade 6-8)	-0.061	0.033	0.164^{*}	0.066

Table 10 continued

Parameter		Schools with TFI Tier 1 scores			
		Elementary		Secondary	
		Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Model 1: Interaction effect models					
Square root of major ODRs per 100 students per day	Prior year outcome (e)	0.945**	0.052	0.945**	0.052
	TFI T1	-0.001	0.001	0.001	0.001
	Year 1 to 4 in implementing SWPBIS (e)	0.028*	0.011	0.028*	0.011
	TFI T1 by Year 1 to 4	-0.001	0.001	0.001	0.002
	Urbanicity (e) - 1 = the most urban	-0.003*	0.002	-0.003*	0.002
Proportions of students with 0 – 1 major ODRs	Prior year outcome (e)	0.851**	0.026	0.851**	0.026
	TFI T1	-0.665	0.901	2.182	1.442
	Year 1 to 4 in implementing SWPBIS (e)	-0.395	0.262	-0.395	0.262
	TFI T1 by Year 1 to 4	-1.309	1.642	1.083	3.886
	% of students eligible for free-/reduced price lunch(e)	-0.012*	0.004	-0.012*	0.004
	Urbanicity (e) - 1 = the most urban	0.108**	0.038	0.108**	0.038
Difference in major ODR rates of between African and non-African students	Prior year outcome (e)	0.578**	0.105	0.578**	0.105
	TFI T1	-0.002	0.002	0.003	0.003
	Year 1 to 4 in implementing SWPBIS (e)	0.007	0.035	0.007	0.035
	TFI T1 by Year 1 to 4	-0.006	0.005	-0.003	0.005
	Middle school (grade 6-8)	-0.058	0.033	0.165*	0.066

Note. $^{*}p < .05.$, $^{**}p < .01$.

Table 10 shows that the significant effect of TFI T1 difference in major ODR rates of between African and non-African students was found only in elementary schools, $b = -0.004$, $p = .044$. This negative association indicates that schools with higher TFI T1 latent scores are likely to have lower levels of disproportionality for African American students in comparison to non-African American students in major ODR per 100 days per student. With an emphasis on moderating effects of grade levels on fidelity effects, Wald tests of equality constraints were conducted in the main effect models. Thus, there were significant between group differences in regression coefficients of TFI T1: $\chi^2(1) = 4.878$, $p = .027$ on square root major ODR rates, $\chi^2(1) = 4.307$, $p = .038$ on proportions of students with 0 – 1 major ODRs. Specifically, implementation of SWPBIS with stronger fidelity functions more effectively in elementary schools than in secondary schools. However, no moderating effect of grade level was found on the racial gap.

With relation to other covariates, SWPBIS implementation within the first 4 years (Year 1 – 4) was significantly positively associated with the square root of ODR rates ($b = 0.028$, $p = .010$) equally in elementary and secondary schools. Inclusion of 4^{th} to 6^{th} graders (middle school) was positively related to the racial gap only in secondary schools ($b = 0.164$, $p = .013$). One demographic covariate was the degree of urbanicity, which was negatively associated with the square root of ODR rates ($b = -0.04$, $p = .040$) and positively with proportions of students with 0 – 1 major ODRs ($b = 0.109$, $p = .004$) equally in elementary and secondary schools. Also, the percent of students eligible for free- or reduced-price lunch was negatively associated with proportions of students with 0 – 1 major ODR ($b = -0.011$, $p = .011$).

Results of Sensitivity Test. The main effect models above were run with the selected sample of schools with TFI T1, T2, and T3 scores. Table 11 indicated that all those models showed a reasonable model fit. As indicated by Table 12, the sensitivity test of TFI T1 showed a significant effect of TFI T1 in elementary schools (b = -0.001, p = .037), in which elementary schools with higher TFI T1 scores would have significantly lower ODR rates. However, the effect of TFI T1 on the racial gap was not statistically significant in the sensitivity test (estimated for both schools without grouping due to the limited sample size of secondary schools).

Table 11.

Model Adequacy of Main Effect Models from Schools with TFI T1, T2, and T3

Model fit indices	Square root of major ODRs per 100 per day	Proportions (%) of students with 0 – 1 major ORs	Racial gap in major ODRs per 100 per day
Chi-Square Test			
Df	21	36	8
Value	29.638	43.029	16.045^{*}
CFI	.998	.995	.977
TLI	.997	.994	.960
SRMR	.058	.053	.051
RMSEA value	.032	.022	.047
RMSEA 90% CI upper bound	.057	.044	.081

Note. $^{*}p < .05$., $^{**}p < .01$.

Table 12

Associations between the TFI and school-wide behavioral outcomes for schools with TFI T1, T2, and T3

		TFI T1			
		Elementary		Secondary	
Outcome	Parameters	Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Square root of major ODRs per 100 students per day	Prior year outcome	0.893**	0.039	1.125**	0.104
	TFI T1	-0.001*	0.001	0.000	0.001
	Year 1 to 4 in implementing SWPBIS	-0.010	0.016	0.070*	0.030
Proportions (%) of students with 0 – 1 major ODRs	Prior year outcome (e)	0.879**	0.028	0.879**	0.028
	TFI T1	0.025	0.016	0.018	0.022
	Year 1 to 4 in implementing SWPBIS	-0.311	0.375	-0.311	0.375
	Urbanicity	0.068	0.058	0.274**	0.091
	% of students eligible for FRL	-0.001	0.007	-0.046**	0.015
Racial gap in major ODRs per 100 per day	Prior year outcome		0.534**		0.129
	TFI T1		-0.004		0.002
	Year 1 to 4 in implementing SWPBIS		-0.021		0.052

Note. $^{*}p < .05.$, $^{**}p < .01.$

Summary. As indicated by Figure 9 to 11, this study found from elementary and secondary schools with TFI T1 scores in 2016-17 and SWIS outcome data consecutively in 2015-16 and 2016-17 that TFI T1 was associated with all behavioral outcomes in the hypothesized directions, but not statistically significantly (except for disproportionality in elementary schools). Particularly in elementary schools, higher TFI T1 scores were significantly associated with lower racial differences in major ODRs. Although the moderating effects of 1 to 4 years of implementing SWPBIS were not found, significant differences between elementary and secondary schools were found on the effects of TFI T1 on major ODR rates and proportions of students with 0 to 1 ODR rates. Further, the sensitivity test with the selected sample of schools with TFI T1, T2, and T3 were shown in Figure 12 to 14, indicating that TFI T1 scores were negatively predictive of major ODR rates.

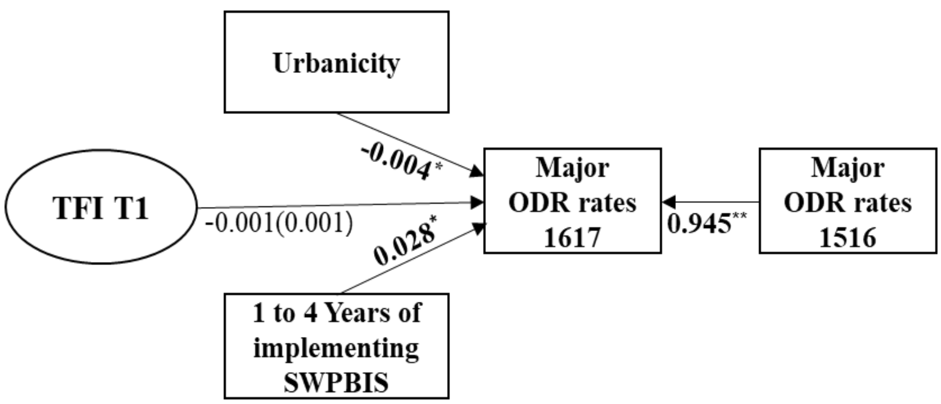

Figure 9. The estimation of association between TFI T1 and square root of major ODRs per 100 students per day. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p$ < .05., $^{**}p$ < .01.) were bolded.

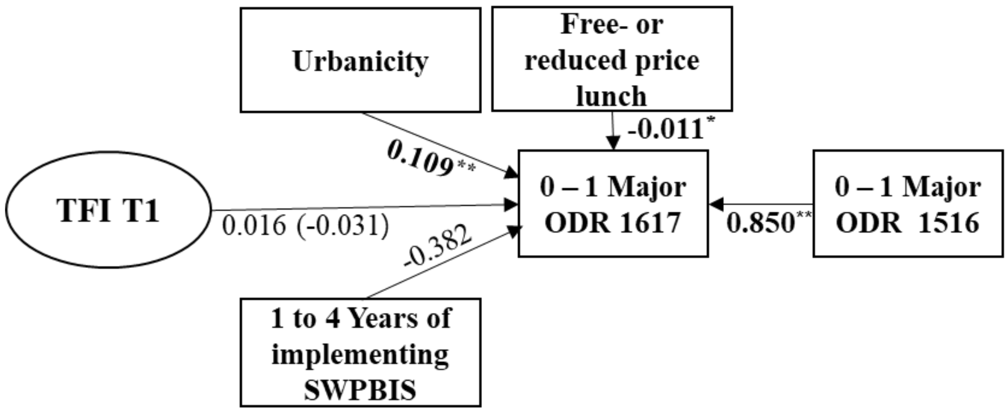

Figure 10. The estimation of association between TFI T1 and arcsine of proportions of students with 0 to 1 major ODRs. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

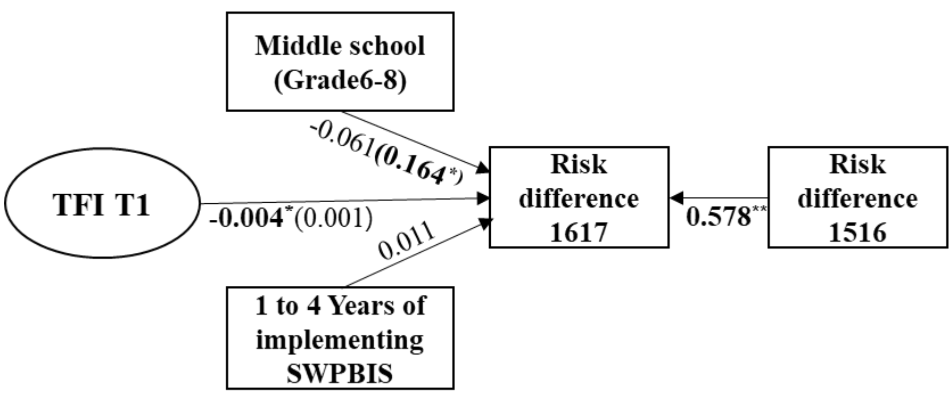

Figure 11. The estimation of association between TFI T1 and risk difference of major ODRs per 100 students per day. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

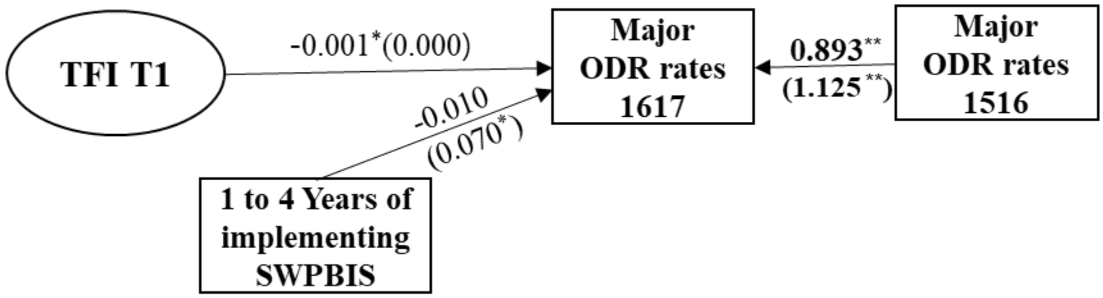

Figure 12. The sensitivity test of the association between TFI T1 and square root of major ODRs per 100 students per day for schools with all TFI T1, T2, and T3. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

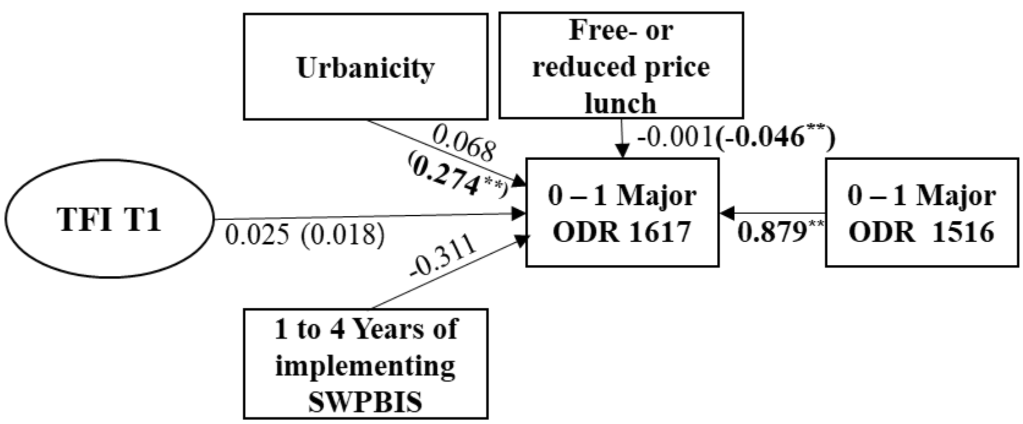

Figure 13. The sensitivity test of the association between TFI T1 and arcsine of proportions of students with 0 to 1 major ODR for schools with all TFI T1, T2, and T3. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

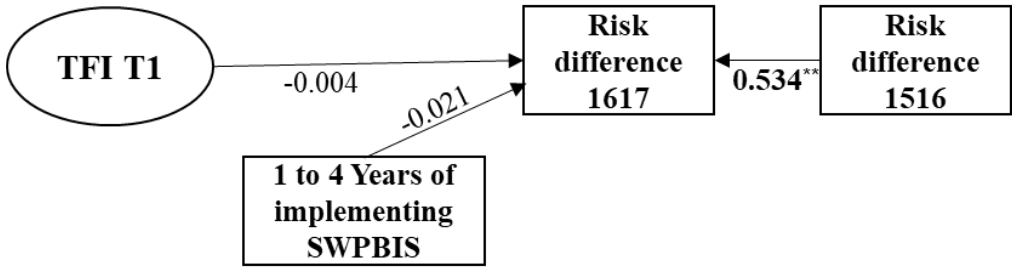

Figure 14. The sensitivity test of the association between TFI T1 and risk difference of major ODRs per 100 students per day for schools with all TFI T1, T2, and T3. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

Research Question 2

Preliminary Analyses. Table 13 indicated that all four invariance models showed reasonable fits to the data based on multiple fit indices. Then, the most parsimonious model, the strictest residual invariance model, was selected for further estimation.

Table 13

Model Fit Statistics for Multi-Group Models of Measurement Invariance

Invariance	Chi-square		CFI	TLI	SRMR	RMSEA	
	Df	Value				Value	90% CI
Configural	64	215.77*	.965	.951	.043	.059	.050 – .068
Metric	71	220.27*	.966	.957	.044	.056	.047 - .064
Scalar	78	242.27*	.963	.957	.047	.056	.048 - .064
Residual	88	257.33*	.961	.961	.053	.053	.046 - .061

$^{*}p < .05$, $^{**}p < .01$.

Table 14 showed all statistically significant factor loadings (.518 to .910), and an acceptable level of composite reliability for three factors (T1 = .74 ~ .80, T2 = .89 ~ .90, T3 = .90 ~ .94). As indicated by Table 14, secondary schools were shown to have significantly lower mean factor scores in TFI Tier 1, Tier 2, and Tier 3.

Table 14

Results of Multi-Group Three Factor Models of Measurement Invariance (*N* = 1,361)

			Elementary			Secondary		
Parameter			*N*	Standardized coefficient	*SE*	*n*	Standardized coefficient	*SE*
T1	Teams	Factor loading	962	0.518**	0.028	399	0.606**	0.032
		Intercept		4.090**	0.129		3.805**	0.125
		Residual		0.732**	0.029		0.633**	0.039
	Implementation	Factor loading	962	0.817**	0.022	399	0.872**	0.017
		Intercept		4.822**	0.194		4.092**	0.167
		Residual		0.332**	0.037		0.239**	0.030
	Evaluation	Factor loading	962	0.775**	0.024	399	0.839**	0.020
		Intercept		4.434**	0.195		3.818**	0.176
		Residual		0.400**	0.037		0.296**	0.034
	Mean			0	0		-0.356**	0.059
	Composite reliability			0.753			0.821	

Table 14 continued

Parameter			Elementary			Secondary		
			N	Standardized coefficient	*SE*	*n*	Standardized coefficient	*SE*
T2	Teams	Factor loading	730	0.831**	0.016	285	0.847**	0.016
		Intercept		4.069**	0.129		2.803**	0.120
		Residual		0.303**	0.026		0.282**	0.028
	Intervention	Factor loading	730	0.903**	0.011	285	0.910**	0.011
		Intercept		2.496**	0.124		2.406**	0.110
		Residual		0.186**	0.020		0.171**	0.021
	Evaluation	Factor loading	730	0.885**	0.012	285	0.894**	0.011
		Intercept		1.718**	0.089		1.653**	0.076
		Residual		0.218**	0.022		0.201**	0.020
	Mean			0	0		-0.364**	0.067
	Composite reliability			0.906			0.915	
T3	Teams	Factor loading	424	0.836**	0.017	177	0.869**	0.016
		Intercept		2.379**	0.119		2.147**	0.094
		Residual		0.301**	0.029		0.244**	0.028
	Resources	Factor loading	424	0.871**	0.016	177	0.898**	0.014
		Intercept		2.061**	0.107		1.845**	0.082
		Residual		0.242**	0.027		0.194**	0.025
	Student Support	Factor loading	424	0.836**	0.020	177	0.869**	0.017
		Intercept		1.937**	0.108		1.748**	0.085
		Residual		0.301**	0.033		0.245**	0.029

Table 14 continued

Parameter			Elementary *n*	Elementary Standardized coefficient	Elementary *SE*	Secondary *n*	Secondary Standardized coefficient	Secondary *SE*
T3	Evaluation	Factor loading	424	0.879**	0.015	177	0.905**	0.011
		Intercept		1.612**	0.091		1.441**	0.066
		Residual		0.228**	0.026		0.182**	0.020
	Mean			0	0		-0.476**	0.087
	Composite reliability			0.916			0.935	
Factor Covariances		T1 ↔ T2		0.723**	0.039		0.801**	0.034
		T2 ↔ T3		0.678**	0.052		0.693**	0.056
		T1 ↔ T3		0.549**	0.052		0.548**	0.065

Note. The measurement models between elementary and secondary schools were constrained to be equal in their (unstandardized) factor loadings, and intercepts and residuals of observed subscale indicators. Also, the factor variances were fixed at 1 to identify the model.

$^{*}p < .05.$, $^{**}p < .01.$

Inspection of multicollinearity. Table 15 indicates that there were only significant associations between TFI T2 and either outcome in 2015-2016 across two levels of schools under the three factor models. However, another model with TFI Tier 1 factor only showed significant effects of TFI Tier 1 on both outcomes in elementary schools, which was also found in other TFI Tier 2 or Tier 3 only models. In a recognition of the shared variance among factors in outcomes, the TFI Tier 1 factor alone was used for further structural models.

Table 15.

Comparison of standardized path coefficients between 3-factor and 1-factor model in academic outcomes for schools implementing SWPBIS

		Standardized coefficients			
		3-factor model		1-factor model(T1)	
Outcome	Predictors	Elementary	Secondary	Elementary	Secondary
Reading in 2015-16	T1	0.042	-0.031	0.060**	-0.018
	T2	0.032	0.024		
	T3	-0.013	-0.013		
	Reading in 2014-15	0.930**	0.850**	0.930**	0.850**
Math in 2015-16	T1	0.014	-0.040	0.033*	-0.013
	T2	0.032	0.077		
	T3	-0.013	-0.068		
	Math in 2014-15	0.934**	0.929**	0.934**	0.928**

Note. $^{*}p < .05$, $^{**}p < .01$.

Measurement models. A 1-factor model was specified with varying equality constraints and compared as shown in Table 16. The most restricted model with equality constraints on all residual variances showed an adequate model fit without obvious gaps to other partial invariance models. As indicated by Table 17, factor loadings ranged from 0.531 to 0.873, and composite reliability in either group was greater than .70. Further, the final measurement model was run with the selected sample of 424 elementary and 175 secondary schools with all TFI T1, T2, and T3 scores, showing reasonable fit as well: χ^2 (7) = 15.070, CFI = .981, TLI = .983, SRMR = .094, RMSEA = .062 (CI = .016 ~ .105).

Table 16

Model Fit Statistics for Multi-Group Models of Measurement Invariance

Invariance	Chi-square *df*	Chi-square Value	CFI	TLI	SRMR	RMSEA Value	RMSEA 90% CI
Metric	2	0.565	1.000	1.005	.014	.000	.000 - .052
Scalar	4	13.906*	.989	.984	.028	.060	.028 - .096
Residual	7	20.784*	.985	.987	.060	.054	.028 - .081

Note. Configural invariance model was just identified and cannot be evaluated via Chi-square test. Metric invariance models fixed the factor means at zero in both groups whereas scalar and residual invariance models fixed the factor mean only in secondary schools.

$^{*}p < .05$, $^{**}p < .01$.

Table 17

Results of Revised Measurement Model for TFI T1

Parameter			Elementary *n*	Elementary Standardized coefficient	Elementary *SE*	Secondary *n*	Secondary Standardized coefficient	Secondary *SE*
T1	Teams	Factor loading	962	0.531**	0.029	399	0.619**	0.033
		Intercept		4.099**	0.131		3.798**	0.126
		Residual		0.718**	0.031		0.617**	0.041
	Implementation	Factor loading	962	0.817**	0.025	399	0.873**	0.018
		Intercept		4.825**	0.193		4.092**	0.166
		Residual		0.332**	0.042		0.239**	0.032
	Evaluation	Factor loading	962	0.767**	0.028	399	0.832**	0.025
		Intercept		4.430**	0.194		3.822**	0.177
		Residual		0.412**	0.042		0.307**	0.041
	Mean			0	0		-0.359**	0.059
	Composite reliability			0.754			0.823	

Note. $^{*}p < .05$, $^{**}p < .01$.

Descriptive Statistics. Table 18 indicates that schools showed higher fidelity over increasing tiers. Consistent with Research question 1, Tier 1 subscale scores in implementation were lower than others in team and evaluation both from elementary and secondary schools. In terms of school grade levels, elementary schools showed higher mean fidelity scores than secondary schools across all subscales.

Table 18

Descriptive Statistics of Study Variables

	Elementary			Secondary		
	N	*M or %*	*SD*	*n*	*M or %*	*SD*
TFI Tier 1 Team (%)	962	83.76	19.87	399	78.13	23.03
TFI Tier 1 Implementation (%)	962	81.77	16.71	399	73.23	19.93
TFI Tier 1 Evaluation (%)	962	82.34	18.83	399	78.07	21.93
TFI Tier 1 (%)	962	82.19	15.24	399	75.77	18.45
TFI Tier 2 (%)	730	71.79	26.01	285	63.37	26.93
TFI Tier 3 (%)	424	66.97	26.28	177	51.28	30.93
Mean-centered reading in 1415	962	-0.13	16.04	399	0.32	15.59
Mean-centered reading in 1516	962	-0.28	15.83	399	0.68	15.12
Mean-centered math in 1415	946	1.87	16.66	393	-4.52	15.41
Mean-centered math in 1516	946	2.06	17.20	393	-4.95	14.70
% of schools implementing for 7+ years [1]	962	40.96	49.20	399	34.34	47.54
% of students eligible for free-reduced lunch	962	58.49	24.62	398	53.29	21.20
% of African American students	962	9.18	15.63	399	9.11	15.47

Note. [1]The variable of 7 or longer years implementing was used as a dummy variable in analyses.

With relation to outcome measures, the proportions of students meeting or exceeding standards in reading from elementary schools negatively deviated from the state mean (for all grade levels) on average whereas those from secondary schools were higher from the state mean (for all grade levels) The proportion of students meetings or exceeding standards in math from elementary schools were higher than the state mean (for all grade levels), whereas those from secondary schools were lower than the state mean (for all grade levels).

The cut-off point of Year 7 was selected on the years of implementing SWPBIS based on the visual inspection of fidelity scores and student outcomes across schools with varying years of implementing SWPBIS. As the patterns of TFI T1 percentage scores from the dataset for Question 2 was almost similar with those for Question 1, additional information was not reported.

Academic outcomes across years of implementing SWPBIS. Figure 15 indicates that mean-centered proportions of students meeting or exceeding the state-level academic standards both in reading and math showed an overall increasing pattern from year 1 to year 7, in which there were more up-and-down variations in math than in reading. In recognition of the overall adequate levels of TFI T1 near .80 after the initial growth (during the first 2 years of implementation), the scatterplots between academic outcomes in 2015-2016 (in the vertical axis) and raw total percentage scores of TFI Tier 1 in 2015-2016 (in the horizontal axis) were compared among three groups of schools with 1 to 3, 4 to 6, and 7 or more years of SWPBIS implementation. Although there were no obvious differences across three levels of years implementing SWPBIS in reading (showing unclear but slightly negative patterns), the scatterplots in math showed that positive pattern were clearly shown for schools implementing SWPBIS during 7 or more years. Recognizing that previous studies (Gage et al., 2016) proposed long-term impacts of fidelity of implementation of SWPBIS on academic outcomes, the cut-off point was initially determined at 7 years (dividing years of implementing SWPBIS into 1 to 6 and 7 or more).

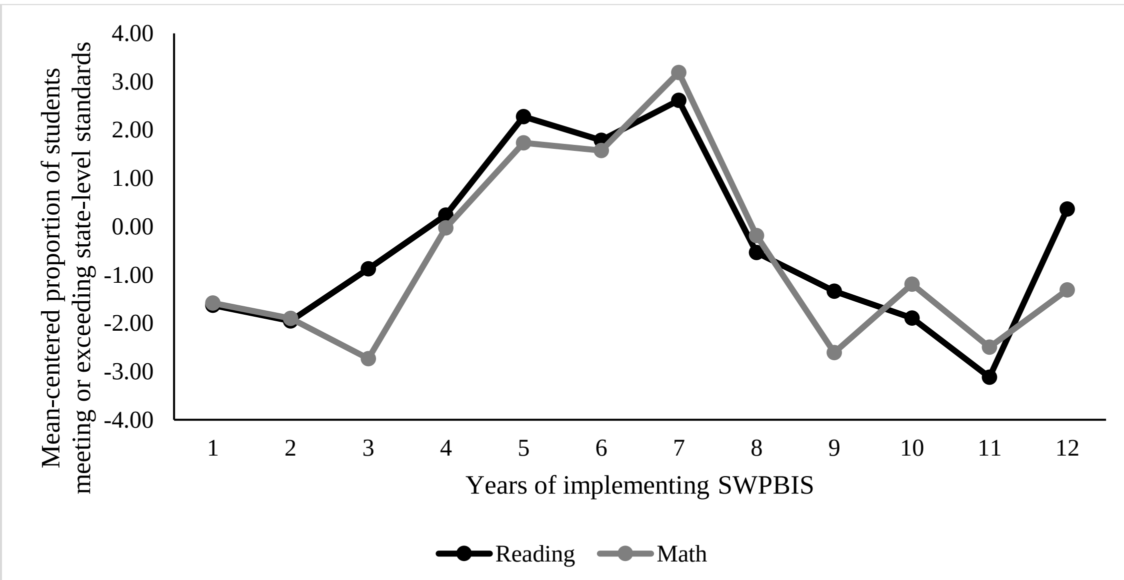

Figure 15. Mean-centered proportions of students meeting or exceeding state-level standards in reading and math test scores. Schools implementing for 12 or more years were not included due to small sample sizes (n = 2 ~ 21).

Results of Structural Models. As shown in Table 19, the final main effect models showed adequate model fit. Interaction effect models extended from the main effect models (as a null model) were evaluated via an Omnibus loglikelihood ratio Chi-square test. As a result, adding interaction terms did not generate significant improvements in either reading or math. Overall, both main effects and interaction effect models can be considered to have adequate fit to the dataset although according to AIC and BIC, although the interaction effect model showed better fit for math achievement, whereas the main-effect model did for reading achievement. As there was still a need for further examinations of interaction effect models due to the possible between group variations under the multi-group models, both models were estimated in sequence and compared to more precisely interpret the interaction effect models.

Table 19

Model Fit Statistics of Final Structural Models for Academic Outcomes

	Reading		Math	
Model fit indices	Main effect model	Interaction effect model	Main effect model	Interaction effect model
Loglikelihood				
Free parameter	19	21	20	22
H0 Scaling correction factor	2.6304	2.4760	1.9546	1.9025
H0 Value	-21738.921	-21738.121	-21268.056	-21264.248
CFI	.975		.985	
TLI	.973		.983	
SRMR	.058		.057	
RMSEA value	.048		.047	
RMSEA 90% CI upper bound	.061		.060	
AIC	43515.843	43518.242	42576.111	42572.496
BIC	43614.932	43627.762	42680.090	42686.873

Main effect models. As reported in Table 20, significant positive effects of TFI Tier 1 were found from elementary schools only in reading (*b* = 0.064, *SE* = 0.022, *p* = .004). In addition, the moderated fidelity effects by grade levels were tested via the Wald test of parameter constraints, showing a significant moderating effect only in reading, $\chi^2(1) = 5.771$, *p* = .016. Related to demographic covariates, schools with larger proportions of students eligible for FRL had significantly lower achievement in both elementary and secondary schools in both reading (*b* = - 0.092, *SE* q= 0.016, *p* < .001) and math (*b* = - 0.078, *SE* = 0.012, *p* < .001).

Table 20

Results of Final Structural Models for Academic Outcomes

Parameter		Elementary		Secondary	
		Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Model 0: Main effect models					
Reading	Mean-centered reading in 1415 (e)	0.796**	0.027	0.796**	0.027
	TFI T1	0.064**	0.022	-0.044	0.038
	Implementing for 7 or more years (e)	0.423	0.420	0.423	0.420
	% of students eligible for free-/reduced price lunch (e)	-0.092**	0.016	-0.092**	0.016
Math	Mean-centered math in 1415	0.875**	0.016	0.825**	0.027
	TFI T1	0.037	0.022	-0.020	0.027
	Implementing for 7 or more years (e)	0.690	0.358	0.690	0.358
	% of students eligible for free-/reduced price lunch (e)	-0.078**	0.012	-0.078**	0.012

Table 20 continued

Parameter		Elementary		Secondary	
		Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Model 1: Interaction effect model					
Reading	Mean-centered reading in 1415 (e)	0.797**	0.027	0.797**	0.027
	TFI T1	0.045	0.030	-0.061	0.058
	Implementing for 7 or more years (e)	0.411	0.427	0.411	0.427
	TFI T1 by Years	0.042	0.041	0.043	0.068
	% of students eligible for free-/reduced price lunch (e)	-0.091**	0.016	-0.091**	0.016
Math	Mean-centered math in 1415	0.876	0.016	0.826**	0.027
	TFI T1	-0.013	0.031	-0.031	0.040
	Implementing for 7 or more years (e)	0.652	0.361	0.652	0.361
	TFI T1 by Years	0.109*	0.045	0.028	0.051
	% of students eligible for free-/reduced price lunch (e)	-0.077**	0.012	-0.077**	0.012

Note. Mean scores of TFI T1 were constrained at zero in both groups. Each structural model was tested respectively for individual outcome measure. $^{*}p < .05.$, $^{**}p < .01$.

Interaction effect models. There was a significant positive interaction effect (b = 0.109, SE = 0.045, p = .016) only for elementary schools in math, which indicates that the effect of TFI Tier 1 would become stronger if schools were implementing SWPBIS for 7 or more years. Although there were no significant effects of TFI T1 under the interaction effect models, understanding the association between TFI T1 and academic outcomes may require careful interpretation under the significant interaction effects suggesting varying fidelity effects over years.

To precisely interpret the interaction effects, the means of the mean-centered reading and math outcomes in 2014-2015 and 2015-2016 were plotted for schools with TFI T1 scores at 1 SD above the mean (higher fidelity) and those at 1 SD below the mean (lower fidelity) depending on whether schools had been implementing SWPBIS for 7 or more years, which were depicted in Figure 16 to 18.

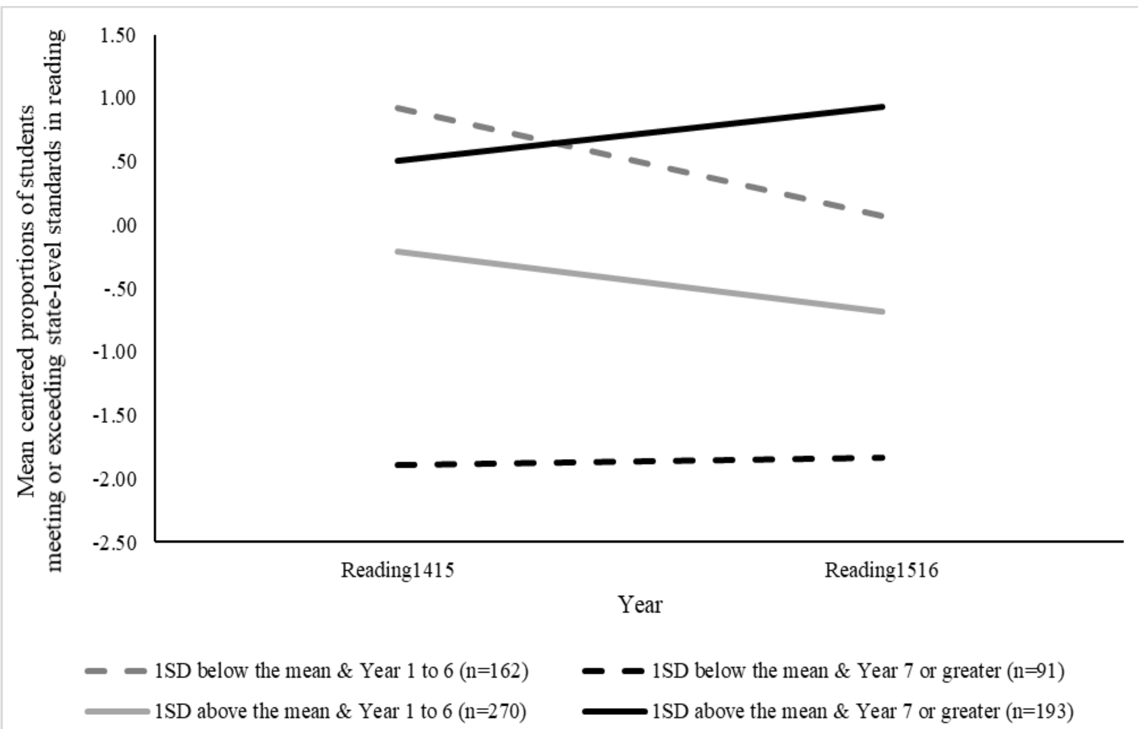

Figure 16. Reading achievement across TFI T1 fidelity and years implementing SWPBIS for elementary schools.

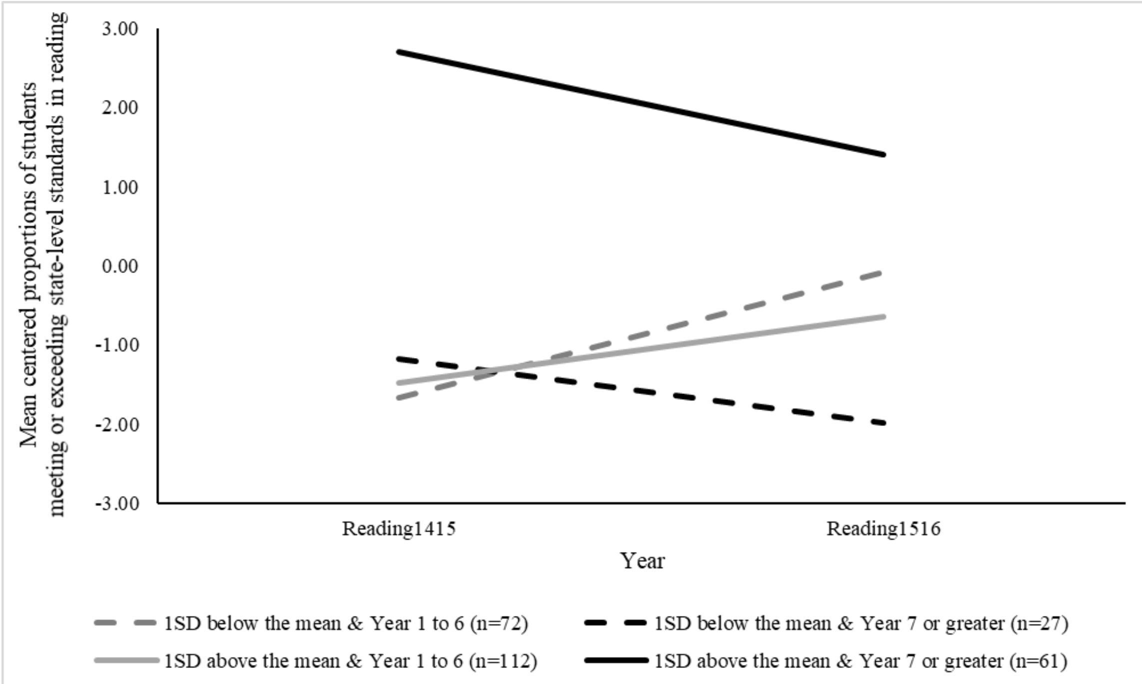

Figure 17. Reading achievement across TFI T1 fidelity and years implementing SWPBIS for secondary schools.

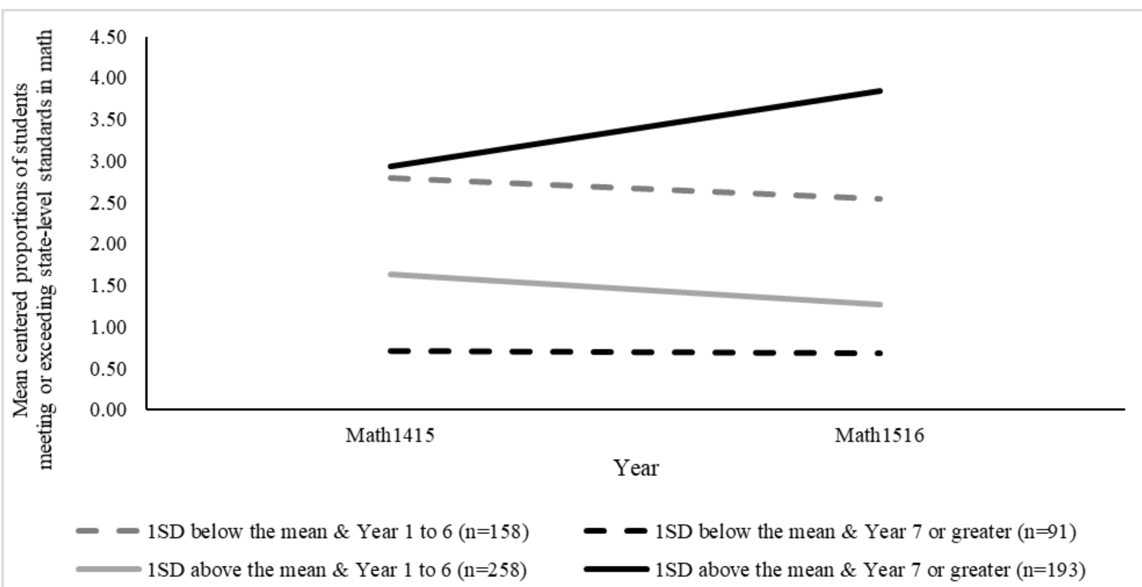

Figure 18. Math achievement across TFI T1 fidelity and years implementing SWPBIS for elementary schools.

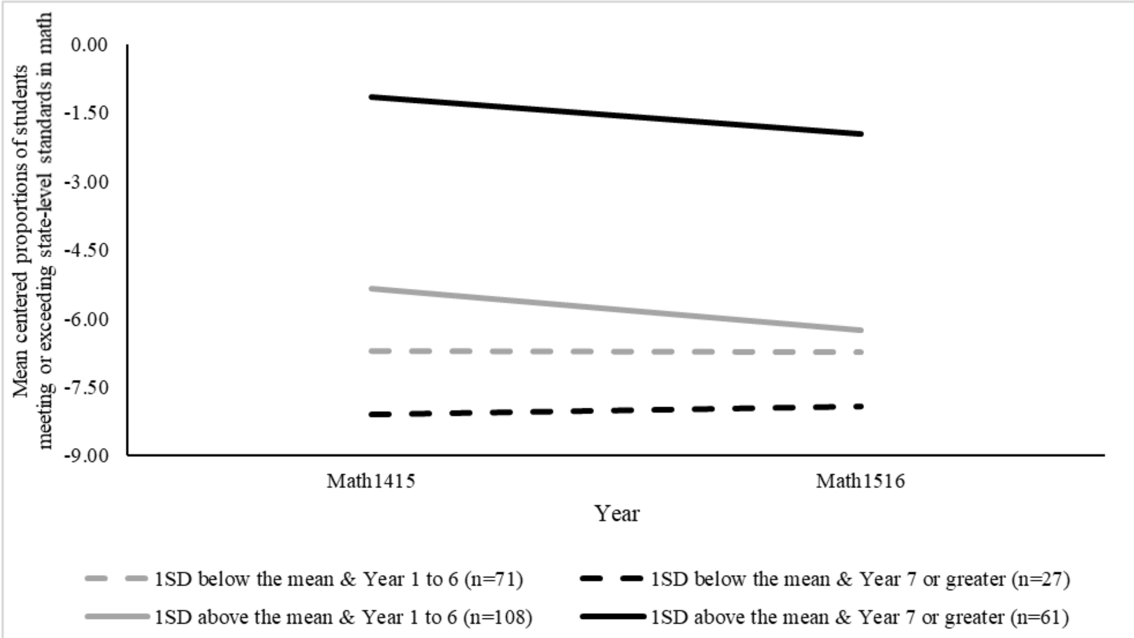

Figure 19. Math achievement across TFI T1 fidelity and years implementing SWPBIS for secondary schools.

In Figure 16, schools with higher fidelity showed higher levels of reading achievement than those with lower fidelity during 7 or more years of SWPBIS implementation. However, schools with fewer than 7 years of SWPBIS implementation showed a somewhat different pattern where schools with lower fidelity (than those with higher fidelity) showed higher levels of reading achievement despite the decreased gap in 2015-2016. For secondary schools shown in Figure 17, schools with higher fidelity (than those with lower fidelity) showed higher levels of reading achievement when they were in 7 or more years of SWPBIS implementation. For schools with 1 to 6 years of implementing SWPBIS, secondary schools with lower fidelity (than those with higher fidelity) exhibited high levels of achievement in reading, and those gaps were larger in 2015-2016. None of these differences across groups were statistically significant.

With regards to math achievement, Figure 18 indicates that elementary schools with higher fidelity (than those with lower fidelity) showed higher levels of math achievement when they were implementing for 7 or more years. However, such patterns flipped during the initial six years of implementation, in which schools with lower fidelity outperformed their counterparts with higher fidelity. In similar ways, Figure 19 indicates that when schools were implementing SWPBIS for fewer than 7 years, secondary schools with higher fidelity stayed lower in their math achievement than those with lower fidelity. When implementing SWPBIS for 7 or more years, secondary schools with higher fidelity displayed higher levels of math achievement than their counterparts with lower fidelity, and those differences only were statistically significant, $t(86) = -2.333$, $p = .022$ in 2014-2015 and $t(86) = -.2.011$, $p = .047$ in 2015-2016.

Results of Sensitivity Test. Due to the insufficient sample size for each group, only the main effect models were run with the selected sample of schools with TFI T1, T2, and T3 scores. Table 21 indicated that both models for reading and math achievement showed a reasonable fit to the dataset. As shown in Table 22, the sensitivity test of TFI T1 showed significant effects of TFI T1 both on reading ($b = 0.071$, $p = .016$) and math ($b = 0.087$, $p = .013$) in elementary schools, in which elementary schools with higher TFI T1 scores would have significantly higher proportions of students meeting or exceeding the state standards in reading or math. Also, 7 or more years of SWPBIS implementation was significantly associated with reading ($b = 1.359$, $p = .031$) only in elementary schools and with math achievement equally across elementary and secondary schools ($b = 0.949$, $p = .044$).

Table 21

Model Adequacy of Main Effect Models from Schools with TFI T1, T2, and T3 for Academic Outcomes

Model fit indices	Reading	Math
Chi-Square Test		
Df	32	32
Value	44.492	54.719*
CFI	.990	.983
TLI	.988	.981
SRMR	.069	.073
RMSEA value	.036	.049
RMSEA 90% CI upper bound	.060	.071

Note. The means of the factor scores each group were constrained at zero.

Table 22

Associations between Academic Achievement and TFI Scores from Schools with TFI T1, T2, and T3

		TFI T1			
		Elementary		Secondary	
Outcome	Parameters	Unstandardized coefficient	*SE*	Unstandardized coefficient	*SE*
Reading	Prior year outcome (e)	0.785**	0.032	0.785**	0.032
	TFI T1	0.071*	0.030	-0.049	0.039
	Implementing for 7 or more years	1.359**	0.630	-0.867	0.956
	% of students eligible for FRL (e)	-0.116**	0.021	-0.116**	0.021
Math	Prior year outcome	0.850**	0.025	0.770**	0.043
	TFI T1	0.087*	0.035	-0.044	0.035
	Implementing for 7 or more years (e)	0.949*	0.470	0.949*	0.470
	% of students eligible for FRL (e)	-0.100**	0.019	-0.100**	0.019

Note. $^{*}p < .05.$, $^{**}p < .01$.

Summary. As shown in Figure 16 to 19, both elementary and secondary schools with higher fidelity outperformed their counterparts with lower fidelity in reading and math of the 2015–16 school year when they had implemented SWPBIS for 7 or more years, whereas inconsistent patterns were found from schools implementing SWPBIS for 6 or fewer years. Comparisons of the main and interaction effects led to the conclusion that the moderating effects of 7 or more years of implementing SWPBIS on fidelity effects on math achievement were empirically supported by this study (depicted in Figure 21); and in addition to the positive main effect of TFI T1 on reading achievement, moderating effects of grade levels on fidelity effects were found on reading achievement (depicted in Figure 22).

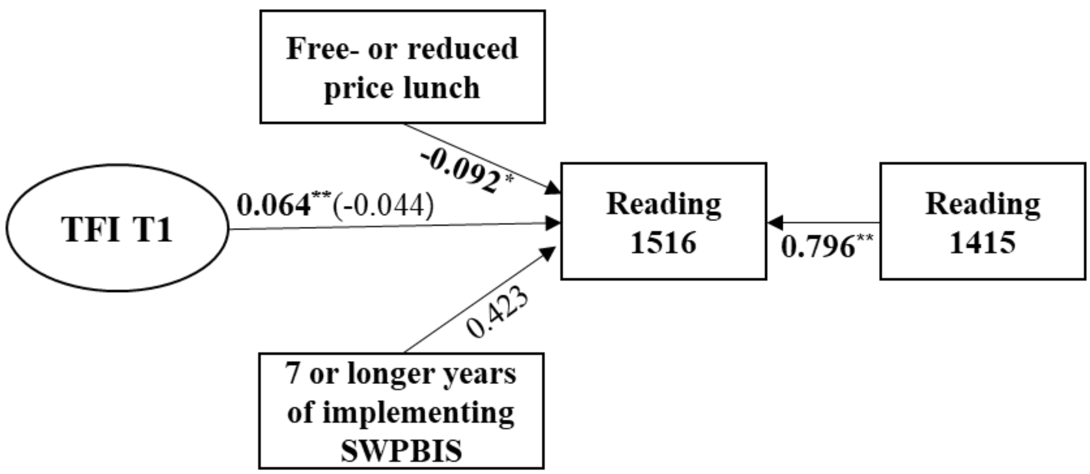

Figure 20. The estimation of association between TFI T1 and state-mean centered proportions of students meeting or exceeding state level benchmark in reading test scores. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

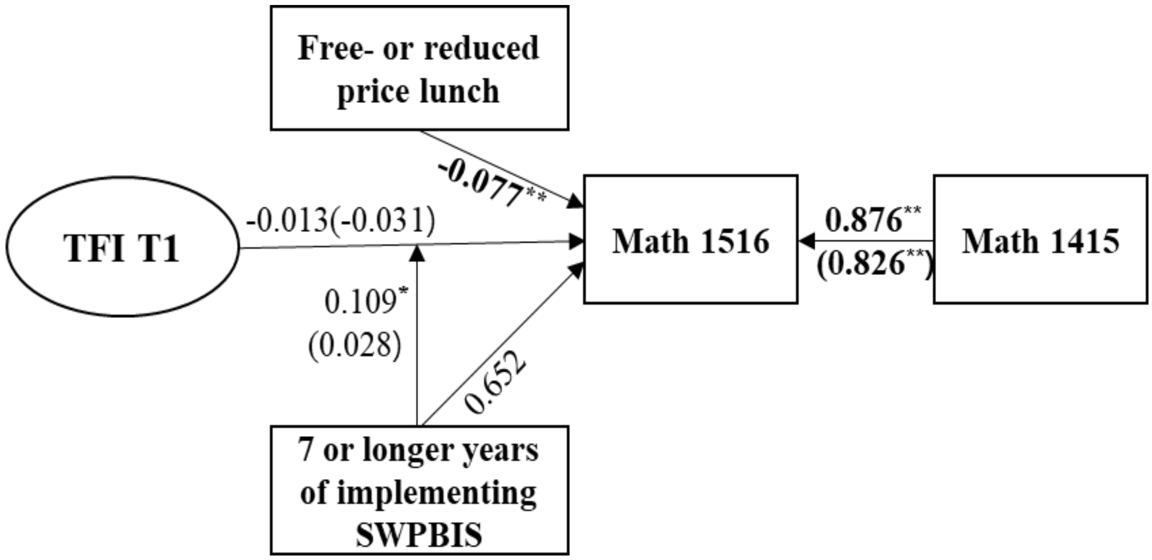

Figure 21. The estimation of association between TFI T1 and state-mean centered proportions of students meeting or exceeding state level benchmark in math test scores. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p$ < .05., $^{**}p$ < .01.) were bolded.

Nevertheless, the non-significant interaction term of fidelity by 7 or more years implementing SWPBIS may not lead to the conclusion of lack of moderating effect of years implementing on fidelity effects because the descriptive exploration of the selected data (within ±1*SD*) visually depicted the varying patterns of reading achievement across the mixture of different years implementing and fidelity. Altogether, the results of the main effect model to support a significantly positive academic effect of TFI T1 was more suitable for reading achievement data in this study, depicted by Figure 16.

In the end, the results of sensitivity tests with the selected sample of schools with TFI T1, T2, and T3 were shown in Figure 22 and 23. Under the main effect models without interaction terms due to insufficient sample size, the results indicated the positive association between TFI T1 and the mean-centered proportions of students meeting or exceeding state-wide benchmarks in reading and math test scores from elementary schools. Further, reading achievement was positively associated with 7 or longer years of implementing SWPBIS only in elementary schools, and math achievement was so in both elementary and secondary schools. Overall, this study supported a positive academic effect of TFI T1 in elementary schools implementing SWPBIS at all three tiers if years of implementing SWPBIS were being held constant across schools.

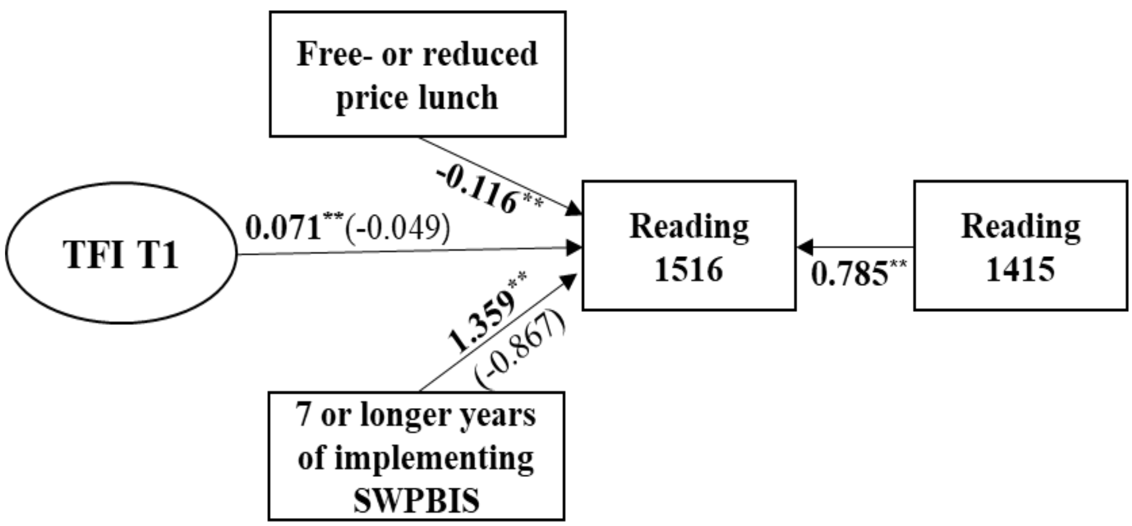

Figure 22. The sensitivity test of the association between TFI T1 and state-mean centered proportions of students meeting or exceeding state level benchmark in reading test scores for schools with all TFI T1, T2, and T3 scores. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

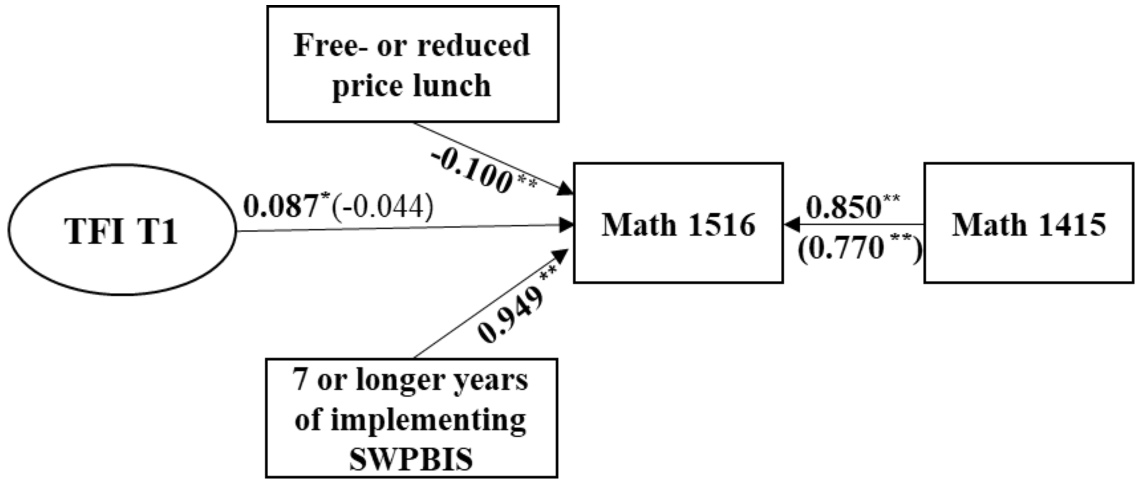

Figure 23. The sensitivity test of the association between TFI T1 and state-mean centered proportions of students meeting or exceeding state level benchmark in math test scores for schools with all TFI T1, T2, and T3 scores. If there were significant differences in regression coefficients of covariates between elementary and secondary schools, coefficients for secondary schools were reported in parentheses. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

Research Question 3

Preliminary Analyses. The three-factor model was initially specified across two groups, generating negative residuals (called Heywood cases, Chen, Bollen, Paxton, Curran, & Kirby, 2001) in the team subscales of all three tiers across two groups, which was different from other models in this study. In response to the relatively small sample size for Question 3 (e.g., 114 secondary schools, 456 elementary schools), grade level was included as a covariate instead of the multi-group analyses. However, the same issue occurred even without grouping. Possible causes of Heywood cases might relate to empirical under-identification (Chen et al., 2001; Rindskopf, 1984), indicating that the model might not be identified only with the sample introducing or implementing CICO in this study. Instead of using subscale scores, alternative specification was made with 13 items to estimate three latent variables corresponding to subscale scores of TFI T2. To best suit the 3-point Likert scale, the item-based model was estimated via estimator = WLSMV. Table 23 indicated that overall the model showed an adequate fit to the data. As reported by Table 24, the factor loadings were statistically significant, ranging from 0.647 to .868. The composite reliability ranged from .841 to .890.

Table 23

Model Fit Statistics for Measurement Model

Model	Chi-square		CFI	TLI	WRMR	RMSEA	
	df	Value				Value	90% CI
TFI T2	62	126.87**	.984	.980	.828	.043	.032 – .053

Note. $^{*}p < .05$., $^{**}p < .01$.

***Inspection of multicollinearity*.** Comparisons of the results of between 3-factor and 1-factor models found that dropping two factors produced differences in coefficients as shown in Table 25. Specifically, the effect of team subscale scores was statistically significant in the 1-factor model for the logit of the proportions of students with CICO daily point data whereas not significant and even negative in the 3-factor model. Due to the detected shared variances in outcomes associated with three factors, only the TFI Tier 2 factor was included in the structural models.

Table 24

Results of Measurement Model of TFI Tier 2 (*N* = 570)

Parameter			Standardized coefficient	*SE*
Team	Item1	Factor loading	0.711**	0.055
		Thresholds_1	-2.107**	0.127
		Thresholds_2	-1.136**	0.076
	Item2	Factor loading	0.738**	0.034
		Thresholds_1	-1.708**	0.090
		Thresholds_2	-0.576**	0.060
	Item3	Factor loading	0.868**	0.034
		Thresholds_1	-1.968**	0.111
		Thresholds_2	-0.792**	0.070
	Item4	Factor loading	0.694**	0.038
		Thresholds_1	-1.708**	0.103
		Thresholds_2	-0.392**	0.068
	Composite reliability		0.841	
Interventions	Item5	Factor loading	0.762**	0.032
		Thresholds_1	-1.999**	0.123
		Thresholds_2	-0.602**	0.070
	Item6	Factor loading	0.825**	0.025
		Thresholds_1	-1.968**	0.115
		Thresholds_2	-0.591**	0.068
	Item7	Factor loading	0.857**	0.024
		Thresholds_1	-1.747**	0.102
		Thresholds_2	-0.591**	0.070

Table 24 continued

Parameter			Standardized coefficient	*SE*
Interventions	Item8	Factor loading	0.776**	0.056
		Thresholds_1	-2.196**	0.138
		Thresholds_2	-1.462**	0.094
	Item9	Factor loading	0.704**	0.032
		Thresholds_1	-1.671**	0.100
		Thresholds_2	-0.062	0.060
	Composite reliability		0.890	
Evaluation	Item10	Factor loading	0.818**	0.029
		Thresholds_1	-1.573**	0.090
		Thresholds_2	-0.583**	0.060
	Item11	Factor loading	0.828**	0.030
		Thresholds_1	-1.811**	0.103
		Thresholds_2	-0.524**	0.060
	Item12	Factor loading	0.647**	0.041
		Thresholds_1	-0.982**	0.074
		Thresholds_2	0.163*	0.064
	Item13	Factor loading	0.739**	0.034
		Thresholds_1	-1.400**	0.087
		Thresholds_2	-0.445**	0.068
	Composite reliability		0.845	
Factor Covariances	Teams ↔ Interventions		0.949**	0.023
	Teams ↔ Evaluations		0.856**	0.030
	Interventions ↔ Evaluations		0.929**	0.024

Note. For model identification, variances were fixed at 1, and the factor means were fixed at 0 as well.

$^{*}p < .05.$, $^{**}p < .01.$

Table 25.

Comparison of standardized path coefficients between 3-factor and 1-factor model for schools implementing CICO

		Standardized coefficients	
Outcome	Predictors	3-factor model	1-factor model
Logit of the proportions of students with CICO daily point data	Team	0.301	0.070*
	Intervention	-0.581	
	Evaluation	0.370	
	The prior outcome	0.718**	0.718**
Percent of students with 80% or above goal accomplishments	Team	0.382	-0.024
	Intervention	-0.469	
	Evaluation	-0.075	
	The prior outcome	0.359**	0.359**
Square root of major ODRs per 100 days per student for students enrolled in CICO	Team	0.248	-0.056
	Intervention	-0.381	
	Evaluation	0.099	
	The prior outcome	0.729**	0.743**

Note. $^{*}p < .05.$, $^{**}p < .01$.

Measurement models. A 1-factor model was specified using 13 items for TFI Tier 2. School grade level was included as a covariate. Table 26 and 27 indicates that the specified measurement model showed adequate fit and reasonable reliability. The factor loadings were all greater than .40, ranging from 0.621 to 0.851. Also, the final measurement model was run with the selected sample of 392 schools (311 elementary and 81 secondary schools) with all TFI T1, T2, and T3 scores, showing a reasonable fit as well: χ^2 (65) = 115.482, CFI = .980, TLI = .976, WRMR = .822, RMSEA = .045 (CI = .031 ~ .058).

Table 26

Model Fit Statistics for Measurement Model

Model	Chi-square		CFI	TLI	WRMR	RMSEA	
	Df	Value				Value	90% CI
TFI T2	65	154.91**	.978	.974	.935	.049	.039 – .059

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

Table 27

Results of Measurement Model of TFI Tier 2 ($N = 570$)

Parameters		Standardized coefficient	*SE*
Item1	Factor loading	0.683**	0.055
	Thresholds_1	-2.107**	0.127
	Thresholds_2	-1.136**	0.075
Item2	Factor loading	0.706**	0.034
	Thresholds_1	-1.708**	0.090
	Thresholds_2	-0.576**	0.060
Item3	Factor loading	0.828**	0.031
	Thresholds_1	-1.968**	0.111
	Thresholds_2	-0.792**	0.070
Item4	Factor loading	0.665**	0.038
	Thresholds_1	-1.708**	0.103
	Thresholds_2	-0.392**	0.068
Item5	Factor loading	0.756**	0.032
	Thresholds_1	-1.999**	0.123
	Thresholds_2	-0.602**	0.070

Table 27 continued

Parameters		Standardized coefficient	*SE*
Item6	Factor loading	0.819**	0.026
	Thresholds_1	-1.968**	0.115
	Thresholds_2	-0.591**	0.068
Item7	Factor loading	0.851**	0.024
	Thresholds_1	-1.747**	0.102
	Thresholds_2	-0.591**	0.070
Item8	Factor loading	0.772**	0.056
	Thresholds_1	-2.196**	0.138
	Thresholds_2	-1.462**	0.094
Item9	Factor loading	0.698**	0.032
	Thresholds_1	-1.671**	0.100
	Thresholds_2	-0.062	0.060
Item10	Factor loading	0.781**	0.029
	Thresholds_1	-1.573**	0.090
	Thresholds_2	-0.586**	0.060
Item11	Factor loading	0.790**	0.030
	Thresholds_1	-1.811**	0.103
	Thresholds_2	-0.524**	0.061
Item12	Factor loading	0.621**	0.041
	Thresholds_1	-0.982**	0.074
	Thresholds_2	0.163*	0.064
Item13	Factor loading	0.703**	0.033
	Thresholds_1	-1.400**	0.087
	Thresholds_2	-0.445**	0.068
Composite reliability		0.942	

$^{*}p < .05$, $^{**}p < .01$.

Descriptive Statistics. Table 28 indicates the means and standard deviations of study variables in Research Question 3. Overall, schools with TFI T2 scores showed strong fidelity scores greater than .80 in either TFI T1 or T2 whereas they reported low scores lower than .70 in TFI T3. Of the TFI Tier 2 subscale scores, the Evaluation subscale scores were lower than other subscale scores. In terms of outcome measures, elementary schools had higher proportions of students receiving CICO, whereas secondary schools had higher proportions of students achieving 80% or more of average daily percent points during the school year. The major ODR rates from students enrolled in CICO were higher in elementary schools than in secondary schools.

Table 28

Descriptive Statistics of Study Variables

	Elementary			Secondary		
	n	*M* or %	*SD*	*n*	*M* or %	*SD*
TFI Tier 2 Team (%)	456	86.73	18.33	114	84.21	18.96
TFI Tier 2 Intervention (%)	456	85.33	17.34	114	83.46	20.72
TFI Tier 2 Evaluation (%)	456	77.17	23.32	114	78.29	23.17
TFI Tier 1 (%)	439	87.87	11.92	112	83.64	12.90
TFI Tier 2 (%)	456	83.25	17.17	114	81.71	18.87
TFI Tier 3 (%)	321	68.77	24.42	81	61.18	27.80

Table 28 continued

	Elementary			Secondary		
	n	*M* or %	*SD*	*n*	*M* or %	*SD*
Logit of proportions of students with CICO daily points in 1516	432	-3.34	0.98	106	-3.54	1.22
Logit of proportions of students with CICO daily points in 1617	432	-3.40	1.02	106	-3.62	1.18
% of students with 80 % or above goal accomplishment in 1516	432	73.97	22.83	107	78.73	25.32
% of students with 80 % or above goal accomplishment in 1617	432	73.75	23.22	107	80.78	22.41
Square root of major ODRs per 100 students per day for students enrolled in CICO in 1516	422	1.65	0.85	103	1.48	0.71
Square root of major ODRs per 100 students per day for students enrolled in CICO in 1617	422	1.64	0.87	103	1.54	0.71
% of schools implementing CICO for 3 or fewer years	456	41.45	49.32	114	33.33	47.34
Mean degrees of urbanicity (1=the most urban ~ 12 = the most rural)	445	7.57	3.07	112	7.72	2.88
% of students eligible for free or reduced lunch	445	54.75	24.09	112	53.22	24.72
% of African American students	452	10.12	16.13	113	10.98	17.75

TFI Tier 2 scores across years of implementing SWPBIS. Figure 24 portrayed the pattern of percent implementation fidelity scores in TFI Tier 2 across years of implementing CICO (indicated number of years using CICO-SWIS) in 2016-17. In this data, only schools with CICO daily points in 2015-16 and 2016-17 were selected for the analyses, and thereby, there were no schools in year 10 with CICO data. The average percentage scores of TFI Tier 2 tended to gradually grow during the initial four years (from year 2 to 5), and after slightly dropping in year 6, rebounded to the highest.

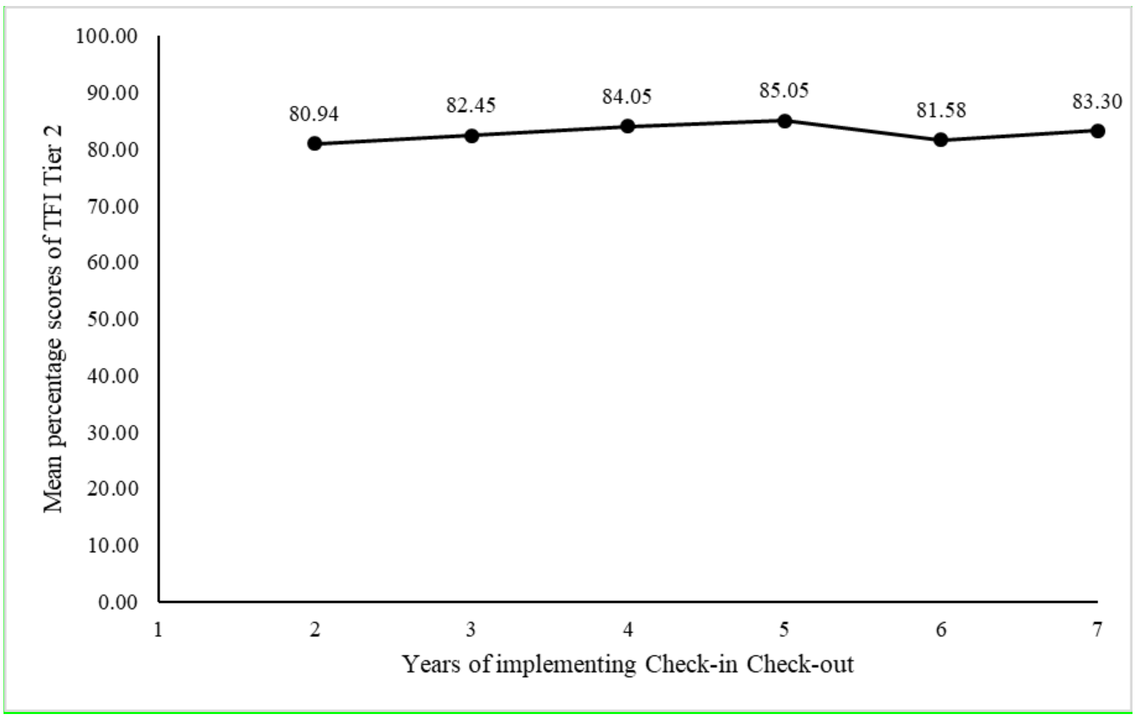

Figure 24. Mean percentage scores of TFI Tier 2 for schools with varying years of implementing Check-in Check-out.

Check-In Check-Out outcomes. Figure 25 and 26 showed that schools with longer years of implementing CICO provided greater proportions of students with CICO and those of students with 80% or greater point goal accomplishment, whereas major ODRs per 100 days per student enrolled in CICO, depicted in Figure 27, were fewer for schools with longer years of implementing CICO. Particularly Figure 26 and 27 involved small fluctuations around year 4 to 6, which is consistent with the pattern of TFI T2 in Figure 24. Overall, schools with more years of CICO implementation tend to have stronger fidelity and better student outcomes. Although the visual inspection of scatterplots suggested that there were not obviously different patterns across schools with varying years of CICO implementation in the extents of associations between fidelity and student outcomes, schools during the initial period might have quicker (or slower) improvement in response to varying fidelity. In this respect, 4 years of CICO implementation (= median) were used as a cut-off to categorize schools into those implementing CICO for 2 to 3 years and more than 4 years.

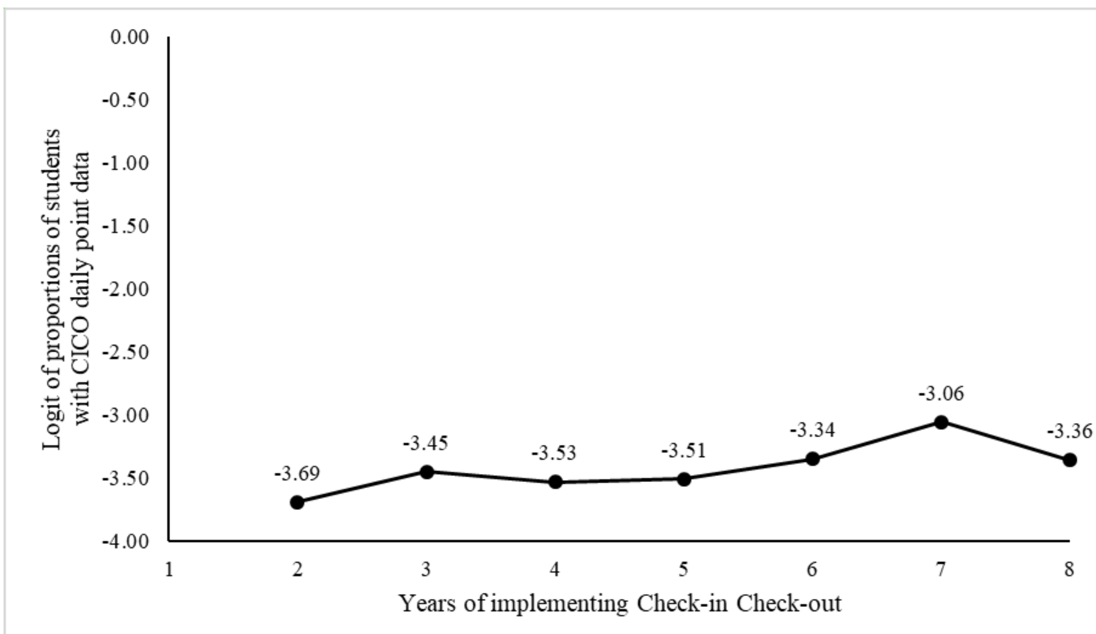

Figure 25. Logit of counts of students with CICO daily points by number of years schools were implementing CICO.

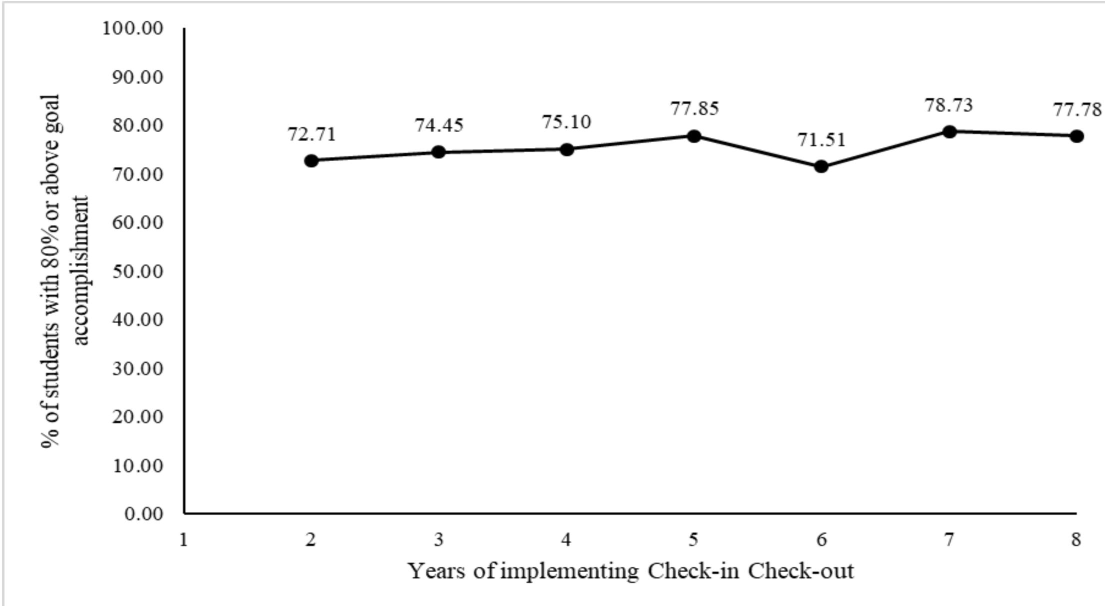

Figure 26. Mean percent of students meeting CICO goals for 80% or more days by number of years schools were implementing CICO.

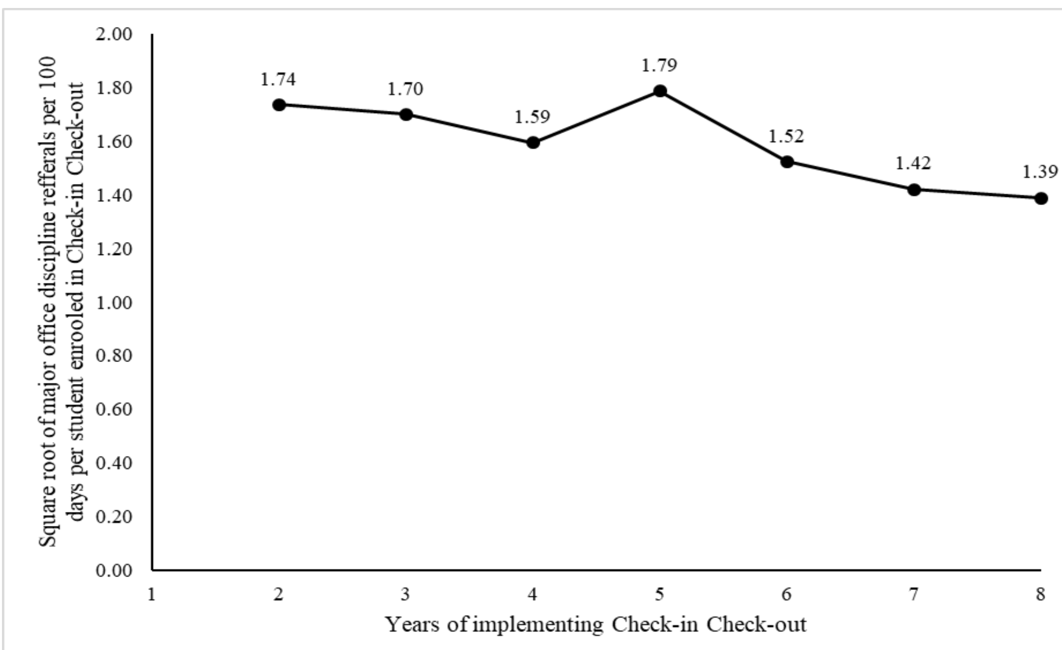

Figure 27. Square root of major ODRs per 100 days per student for students enrolled in CICO by number of years schools were implementing CICO.

Results of Structural Models. Table 29 compared the model fits of between interaction and main effect models. Under WLSMV estimators, the Log likelihood values of main effect modes were computed via using Type = Complex Random. However, due to computational limitation, the Chi-square related fit indices were computed via using Type = Complex. For all outcome measures, the loglikelihood ratio tests indicated that adding interaction terms did not result in statistically significant differences in goodness of fit: $\chi^2(1) = 2.189$, $p = .139$ in the logit of students with CICO daily points, $\chi^2(1) = 0.844$, $p = .358$ in proportions of students with 80% or greater goal accomplishment, and $\chi^2(1) = 0.174$, $p = .677$ in square root of major ODR rates for students enrolled in CICO. As there were no statistically significant interaction terms as indicated by Table 30, the further interpretation focused on the main effect models.

Table 29

Model Fit Statistics of Final Structural Models for CICO Implementers

	Logit of proportions of students with CICO daily points		% of students with 80 % or above goal accomplishment		Square root of major ODRs rates for students enrolled in CICO	
Model fit indices	Main effects	Interaction effects	Main effects	Interaction effects	Main effects	Interaction effects
Loglikelihood						
Free parameter	45	46	46	47	44	45
H0 Scaling correction factor	1.2066	1.2095	1.2192	1.2169	1.2138	1.2132
H0 Value	-4484.737	-4483.270	-6154.294	-6153.819	-4306.316	-4306.213
CFI	.981		.984		.990	
TLI	.978		.982		.989	
WRMR	1.060		1.044		.891	
RMSEA value	.036		.033		.028	
RMSEA 90% CI upper bound	.045		.041		.038	
AIC	9059.473	9058.540	12400.589	12401.638	8700.632	8702.426
BIC	9252.427	9255.782	12596.793	12602.107	8888.886	8894.960

Table 30

Results of Final Structural Models from Schools with TFI T2

Model 0: Main effect models		Unstandardized coefficient	*SE*
Logit of proportions of students with CICO daily points	Prior year outcome	0.726**	0.034
	TFI T2	0.106*	0.053
	Year 2 to 3 in implementing CICO	0.141	0.072
	School size	-0.070*	0.034
% of students with 80 % or above goal accomplishment	Prior year outcome	0.338**	0.040
	TFI T2	-0.785	1.450
	Year 2 to 3 in implementing CICO	0.747	2.058
	% of students eligible for free-/reduced price lunch	-0.100**	0.037
	Elementary School	-5.317*	2.523
Square root of major ODRs rates for students enrolled in CICO	Prior year outcome	0.742**	0.026
	TFI T2	-0.055	0.037
	Year 2 to 3 in implementing CICO	0.018	0.054
Model 1: Interaction effect models			
Logit of proportions of students with CICO daily points	Prior year outcome	0.716**	0.040
	TFI T2	0.082*	0.039
	Year 2 to 3 in implementing CICO	0.141	0.073
	TFI T1 by Year 2 to 3	-0.088	0.062
	School size	-0.065	0.033

Table 30 continued

Model 1: Interaction effect models			
% of students with 80 % or above goal accomplishment	Prior year outcome	0.339**	0.060
	TFI T2	-0.668	0.901
	Year 2 to 3 in implementing CICO	0.812	2.128
	TFI T1 by Year 2 to 3	1.471	1.551
	% of students eligible for free-/reduced price lunch(e)	-0.104*	0.045
	Elementary School	-5.165*	2.280
Square root of major ODRs rates for students enrolled in CICO	Prior year outcome	0.742**	0.044
	TFI T2	-0.033	0.025
	Year 2 to 3 in implementing CICO	0.013	0.056
	TFI T1 by Year 2 to 3	0.018	0.045

Note. $^{*}p < .05.$, $^{**}p < .01$.

Table 30 showed the significant effect of TFI T2 on logit of proportions of students with daily points (enrolled in CICO), $b = 0.106$, $p = .046$. Such a positive effect indicates that schools with higher TFI T2 latent scores are likely to have greater proportions of students enrolled in CICO. With relation to other covariates, school size was negatively associated with logit of proportions of students with daily points (enrolled in CICO), $b = -0.070$, $p = .043$. Also, proportions of students eligible for free or reduced price lunch (*b* = -0.100, *p* = .007) and elementary schools (*b* = -5.317, *p* = .035) were significantly related to % of students with 80% or above goal accomplishments.

Results of Sensitivity Test. The main effect models were run with the selected sample of schools with TFI T1, T2, and T3 scores. Table 31 showed that all those models showed a reasonable fit. As shown in Table 32, the sensitivity test of TFI T2 confirmed the significant effect of TFI T2 on the logit of proportions of students enrolled in CICO with daily points (*b* = 0.157, *p* = .001), in which schools with higher TFI T2 scores would have significantly higher proportions of students enrolled in CICO with daily points.

Table 31

Model Adequacy of Main Effect Models from Schools with TFI T1, T2, and T3 for the CICO Outcomes

Model fit indices	% of students with daily points enrolled in CICO	% of students with 80% or above goal accomplishment	Major ODR rates for students enrolled in CICO
Chi-Square Test			
Df	116	116	103
Value	167.023*	151.277*	128.352*
CFI	.982	.987	.990
TLI	.979	.985	.989
WRMR	0.949	0.886	0.820
RMSEA value	.034	.029	.026
RMSEA 90% CI upper bound	.046	.041	.039

Note. The means of the factor scores each group were constrained at zero.

Table 32

Associations between the TFI and CICO outcomes from Schools with TFI T1, T2, and T3

Model 0: Main effect models		Unstandardized coefficient	*SE*
Logit of proportions of students with CICO daily points	Prior year outcome	0.764^{**}	0.038
	TFI T2	0.157^{**}	0.057
	Year 2 to 3 in implementing CICO	0.156	0.092
	School size	-0.074^{*}	0.037
% of students with 80 % or above goal accomplishment	Prior year outcome	0.453^{**}	0.051
	TFI T2	-1.428	1.515
	Year 2 to 3 in implementing CICO	0.909	2.512
	% of students eligible for free-/reduced price lunch	-0.157^{**}	0.042
Square root of major ODRs rates for students enrolled in CICO	Prior year outcome	0.758^{**}	0.041
	TFI T2	-0.073	0.039
	Year 2 to 3 in implementing CICO	0.035	0.066

Note. $^{*}p < .05.$, $^{**}p < .01$.

Summary. Figure 28 to 30 depicted the result of the final main effect models on the associations between TFI T2 and relevant student outcomes. This study found the positive effect of the TFI T2 on proportions of students with CICO daily points, which is consistent with the result of sensitivity test with the selected sample with TFI T1, T2, and T3. The results of the sensitivity test were depicted by Figure 31 to 33.

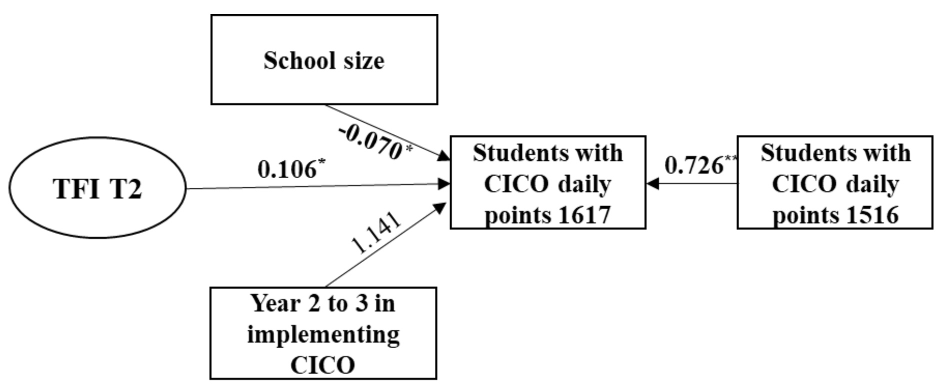

Figure 28. Estimation of association between TFI T2 and arcsine of proportions of students with CICO daily points.

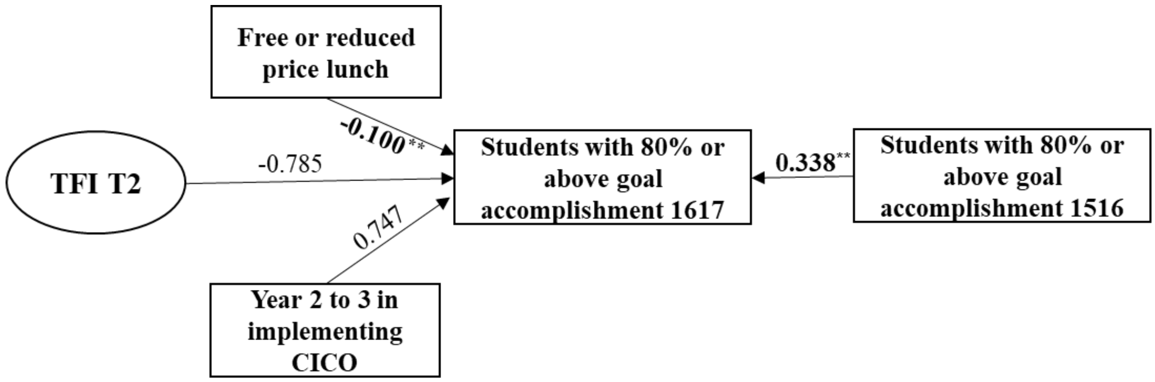

Figure 29. Estimation of association between TFI T2 and proportions of students with 80% or above goal accomplishment. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

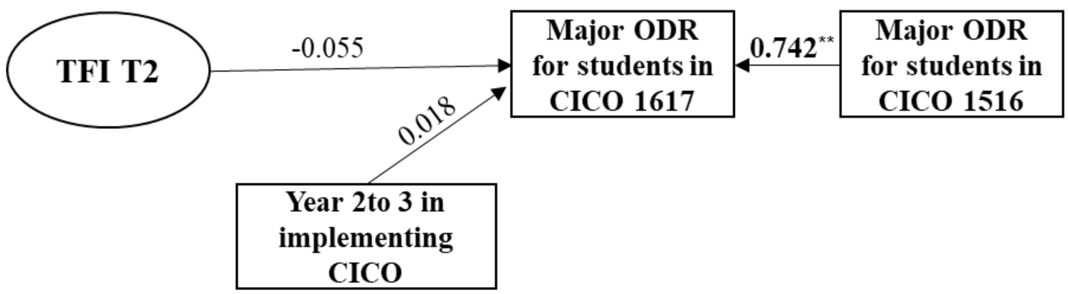

Figure 30. Estimation of association between TFI T2 and square root of major ODR per 100 days per student for students with CICO daily points. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

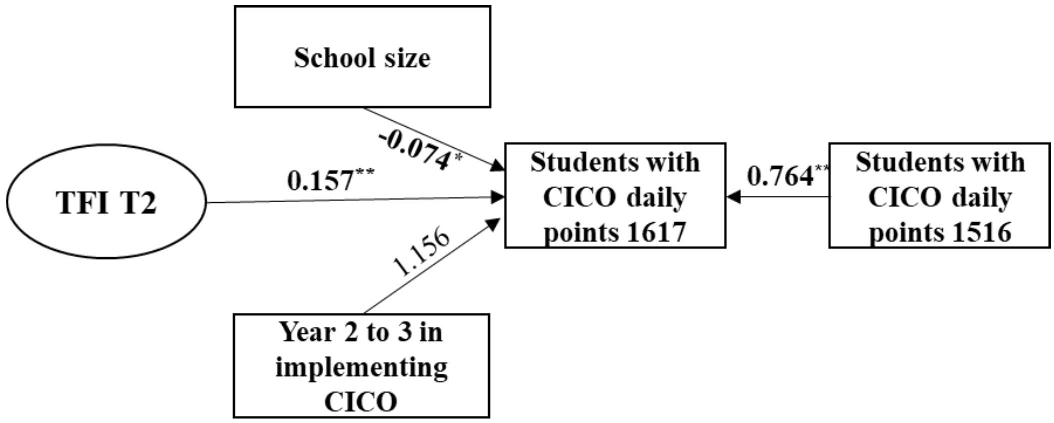

Figure 31. Estimation of association between TFI T2 and arcsine of proportions of students with CICO daily points for schools with all TFI T1, T2, and T3. The statistically significant coefficients ($^{*}p$ < .05., $^{**}p$ < .01.) were bolded.

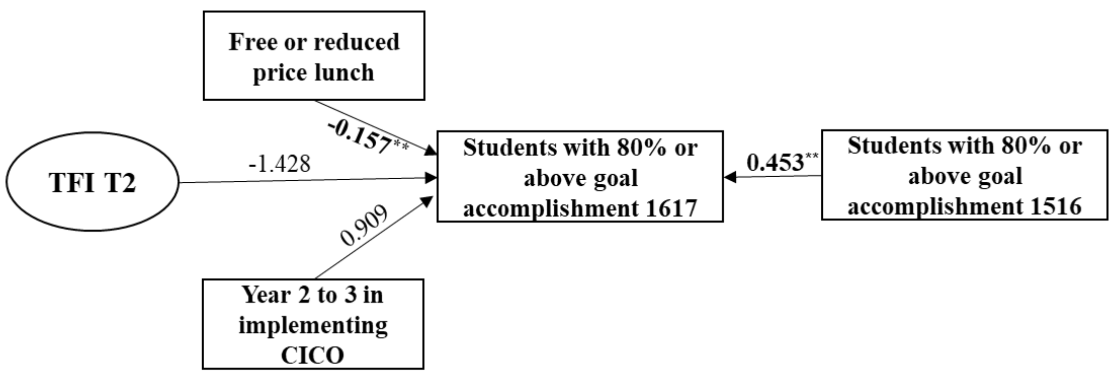

Figure 32. Estimation of association between TFI T2 and proportions of students with 80% or above goal accomplishment for schools with all TFI T1, T2, and T3. The statistically significant coefficients ($^{*}p$ < .05., $^{**}p$ < .01.) were bolded.

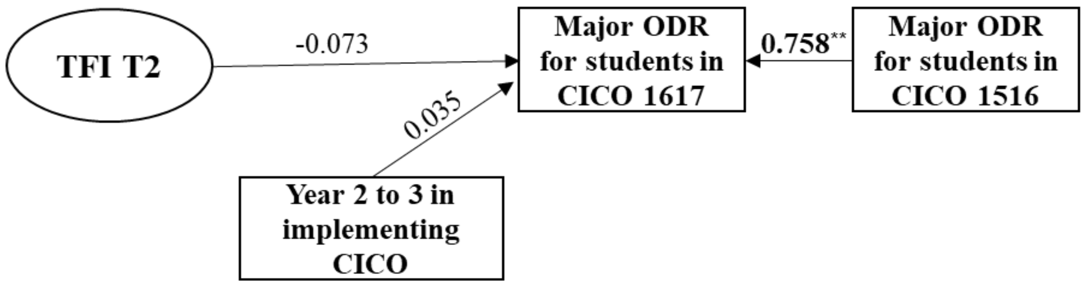

Figure 33. Estimation of association between TFI T2 and square root of major ODR per 100 days per student for students with CICO daily points for schools with all TFI T1, T2, and T3. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

Research Question 4

Preliminary analyses. The evaluation factor each tier was specified via four items in Tier 1 and Tier 2 as shown in Table 33. In a recognition of the model complexity involving numerous covariance paths among study variables, the CFA with ordinal indicators was tested for the entire schools involving all grade level schools. Table 34 showed that all factor loadings were statistically significant, ranging from .692 to .771 for Tier 1 Evaluation and from .760 to .906 for Tier 2 Evaluation. All factor scores showed adequate levels of internal consistency (.82 in T1 Evaluation, .90 in T2 Evaluation).

Table 33

Model Fit Statistics for Evaluation Factors of TFI Tier 1 and Tier 2

Model	Chi-square		CFI	TLI	WR-MR	RMSEA	
	Df	Value				Value	90% CI
T1 Evaluation	2	28.61*	.985	.954	0.978	.075	.052 – .101
T2 Evaluation	2	29.20*	.995	.985	0.931	.088	.061 - .117

Note. $^{*}p < .05.$, $^{**}p < .01.$

Table 34

Results of Evaluation Factor Models of TFI Tier 1 and Tier 2 (N = 2,380)

Parameter			N	Standardized coefficient	SE
Tier1	Item12	Factor loading	2,343	0.696**	0.035
		Thresholds1		-1.995**	0.063
		Thresholds2		-1.254**	0.047
	Item13	Factor loading		0.769**	0.021
		Thresholds1		-1.384**	0.041
		Thresholds2		-0.216**	0.038
	Item14	Factor loading		0.771**	0.026
		Thresholds1		-1.852**	0.059
		Thresholds2		-1.030**	0.043
	Item15	Factor loading		-0.692**	0.024
		Thresholds1		-1.378**	0.050
		Thresholds2		0.016**	0.044
Tier2	Item10	Factor loading	1,774	0.877**	0.011
		Thresholds1		-0.938**	0.048
		Thresholds2		-0.024	0.043
	Item11	Factor loading		0.906**	0.012
		Thresholds1		-1.102**	0.047
		Thresholds2		0.049	0.044
	Item12	Factor loading		0.781**	0.017
		Thresholds1		-0.594**	0.049
		Thresholds2		0.504**	0.043
	Item13	Factor loading		0.760**	0.021
		Thresholds1		-0.829**	0.053
		Thresholds2		0.025	0.053

Note. *p < .05., **p < .01.

Descriptive Statistics. Table 35 indicates the means and standard deviations of study variables in Research Question 4. Overall, TFI Tier 2 evaluation scores were lower than Tier 3 evaluation subscale scores. Elementary schools showed higher mean fidelity scores than secondary schools. The average years of reporting fidelity scores were 4.70 in elementary schools and 4.40 in secondary schools, and the average years of entering

CICO data were 2.74 in elementary schools and 2.67 in secondary schools. Overall, schools reported multiple fidelity scores (2.32 to 2.40 on average) and updated the TFI scores in 1.62 to 1.63 times on average during the school year. With relation to data viewing actions, schools viewed the school summary, average ODRs per day per month, and ODRs by student reports more frequently than other SWIS Reports, whereas schools viewed the student count and average daily points by student more often than other CICO-SWIS Reports.

Table 35

Descriptive Statistics of Study Variables

	Elementary			Secondary		
	N	*M or %*	*SD*	*N*	*M or %*	*SD*
TFI Tier 1 Evaluation (%)	1665	83.05	19.20	644	81.02	20.04
Item12	1665	1.88	0.39	644	1.86	0.41
Item13	1665	1.53	0.64	644	1.43	0.67
Item14	1665	1.82	0.45	644	1.80	0.48
Item15	1665	1.42	0.64	644	1.39	0.65
TFI Tier 2 Evaluation (%)	1287	64.20	30.50	464	58.41	31.35
Item10	1287	1.37	0.75	464	1.25	0.77
Item11	1287	1.40	0.70	464	1.20	0.70
Item12	1287	1.06	0.77	464	0.97	0.75
Item13	1287	1.30	0.78	464	1.24	0.80
Years of SWPBIS implementation[1]	1695	4.70	3.53	651	4.49	3.41
Years of CICO implementation[2]	855	2.74	2.29	254	2.67	2.24
Counts of TFI updates	1695	1.63	0.88	651	1.62	0.85
Number of fidelity measures	1695	2.40	1.29	651	2.32	1.21

Table 35 continued

	Elementary			Secondary		
	N	*M or %*	*SD*	*N*	*M or %*	*SD*
Viewing SWIS-Core Reports - School summary	1695	14.63	13.21	650	13.67	13.24
- Average ODRs per day per month	1695	17.04	26.21	650	18.09	23.68
- ODRs by time	1695	6.03	9.20	650	5.52	8.04
- ODRs by location	1695	10.15	14.30	650	7.38	10.61
- ODRs by day of week	1695	4.76	8.66	650	4.52	7.49
- ODRs by problem behaviors	1695	11.32	15.10	650	16.79	20.99
- ODRs by grade	1695	10.08	13.80	650	10.21	14.74.
- ODRs by student	1695	62.43	80.32	650	62.01	81.50
Viewing SWIS-Additional Reports -Multi-year reports for average ODRs per day per month	1695	10.25	19.57	650	10.93	19.94
-Multi-year reports for ODRs by location	1695	1.97	3.67	650	1.58	3.32
- Multi-year reports for ODRs by problem behaviors	1695	2.59	4.03	650	3.83	5.32
- ODRs by staff	1695	3.44	7.68	650	6.46	14.13
- Suspension/expulsion	1695	5.23	9.92	650	10.70	17.71
- School ethnicity	1695	2.87	6.84	650	3.80	6.89

Table 35 continued

	Elementary			Secondary		
	N	*M or %*	*SD*	*N*	*M or %*	*SD*
Counts of viewing SWIS-CICO Reports - School-wide	737	8.25	12.36	195	8.22	9.49
- Student count	737	66.70	118.00	195	49.96	102.46
- Average daily points by student	737	26.03	43.13	195	21.54	41.37
- Student period	737	11.85	31.64	195	13.67	67.84
- Student single period	737	2.98	10.08	195	2.59	6.86

Note. [1]Years of implementing SWPBIS was indicated by years of reporting fidelity scores. [2]Years of implementing CICO was measured by years of entering SWIS-CICO data.

$^{*}p < .05.$, $^{**}p < .01.$

Results of Structural Models. Two structural models were specified to assess the associations between 1) the TFI T1 Evaluation subscale and generation of SWIS Report data and 2) the TFI T2 Evaluation subscale and generation of SWIS-CICO Report data. Model fits of two models to the data were reasonable as shown in Table 36. Then, two covariance path models between all study variables including subscale scores and other relevant indicators were freely estimated across two grade levels. Results show significant but modest correlations between TFI T1 Evaluation scores and all other variables (ranging from .030 to .262) except counts of viewing SWIS-Additional Reports, ODRs by staff (B = .030, p = .296). Also, TFI T2 Evaluation was significantly but modestly correlated with years of reporting fidelity scores (B = .216, p < .001), years of implementing CICO (*B* = .095, *p* = .035), and counts of viewing the CICO Reports (*B* = -.065 ~ .278) except Student period.

Table 36.

Model Fit of TFI T1 and T2 Evaluation

Model	Chi-square		CFI	TLI	WRMR	RMSEA	
	df	Value				Value	90% CI
T1 Evaluation	53	257.974**	.993	.972	0.942	.040	.035 - .045
T2 Evaluation	29	90.568**	.991	.975	0.825	.030	.023 - .037

Note. $^{*}p$ < .05., $^{**}p$ < .01.

Summary. Figure 34 indicates that the TFI T1 Evaluation subscale was positively correlated with years of implementing SWPBIS, counts of TFI administrations, counts of all fidelity measure administrations, and frequency of viewing SWIS Reports. All correlations with the TFI T1 Evaluation subscale were significant except for counts of viewing the ODRs by Staff Additional Report. Also, Figure 35 shows that the TFI T2 Evaluation subscale was significantly positively associated with years of SWPBIS implementation, years of CICO implementation, and counts of CICO Report – schoolwide, student count, and average daily points by student although positive correlation between the subscale and counts of CICO Report – student period was statistically insignificant. Also, there were negative correlations between TFI T2 Evaluation scores and counts of TFI administrations, counts of all fidelity measure administrations, and counts of viewing CICO Report – student single period, but only the correlation between the subscale and counts of viewing student single period was significant.

Table 37.

Correlations between TFI Tier 1 Evaluation and Counts of Viewing SWIS Report

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1. TFI Tier 1 Evaluation																	
2. Years of SWPBIS implementation[1]	**.140****																
3. Counts of TFI updates	**.142****	-.237**															
4. Number of fidelity measures	**.110****	-.258**	.743**														
SWIS Reports 6. School summary	**.202****	.065*	.076**	.096**													
7. Average ODRs per day per month	**.187****	.057*	.091**	.080**	.249**												
8. ODRs by time	**.236****	.021	.037	.064*	.189**	.456**											
9. ODRs by location	**.231****	.019	.062*	.091**	.230**	.503**	.721**										
10. ODRs by day of week	**.181****	.047	.021	.037	.172**	.356**	.753**	.636**									
11. ODRs by problem behaviors	**.206****	.045	.014	.046	.256**	.478**	.649**	.660**	.577**								
12. ODRs by grade	**.220****	.101**	.006	.044	.269**	.449**	.597**	.600**	.586**	.592**							

Table 37 continued

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
13. ODRs by student	**.262****	.105**	-.030	.012	.326**	.364**	.322**	.369**	.281**	.376**	.382**						
14. Multi-year reports for ODR rates	**.258****	.140**	-.020	-.021	.218**	.318**	.199**	.189**	.215**	.189**	.248**	.265**					
15. Multi-year reports for ODRs by location	**.240****	.102**	-.016	-.003	.223**	.292**	.299**	.357**	.259**	.259**	.275**	.202**	.456**				
16. Multi-year reports for ODRs by problem behaviors	**.205****	.102**	-.032	-.017	.251**	.270**	.249**	.213**	.234**	.372**	.274**	.237**	.441**	.599**			
17. ODRs by staff	**.030**	-.022	.020	.027	.192**	.243**	.170**	.159**	.208**	.259**	.220**	.242**	.280**	.174**	.215**		
18. Suspension/ expulsion	**.070****	.060*	.047	.067*	.226**	.208**	.103**	.120**	.109**	.274**	.168**	.138**	.276**	.181**	.289**	.301**	
19. School ethnicity	**.092***	.106**	.032	-.019	.156**	.181**	.136**	.133**	.130**	.159**	.172**	.136**	.184**	.168**	.231**	.163**	.195**

Note. $^{*}p < .05.$, $^{**}p < .01$.

Table 38.

Correlations between TFI Tier 2 Evaluation and Counts of Viewing CICO-SWIS Report

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1. TFI Tier 2 Evaluation									
2. Years of SWPBIS implementation[1]	**.216****								
3. Years of CICO implementation	**.095***	.517**							
4. Counts of TFI updates	**-.070**	-.237**	-.373**						
5. Number of fidelity measures	**-.021**	-.258**	-.347**	.743**					
CICO Report: 6. School-wide	**.111***	-.116**	-.212**	.335**	.310**				
7. Student count	**.200****	.055	.046	.042	.070	.271**			
8. Average daily points by student	**.278****	.023	.044	.124**	.138**	.396**	.299**		
9. Student period	**.024**	.032	-.002	.044*	.022	.146**	.159**	.255**	
10 Student single period	**-.065***	-.072	-.113**	.165**	.104**	.225**	.198**	.196**	.160**

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

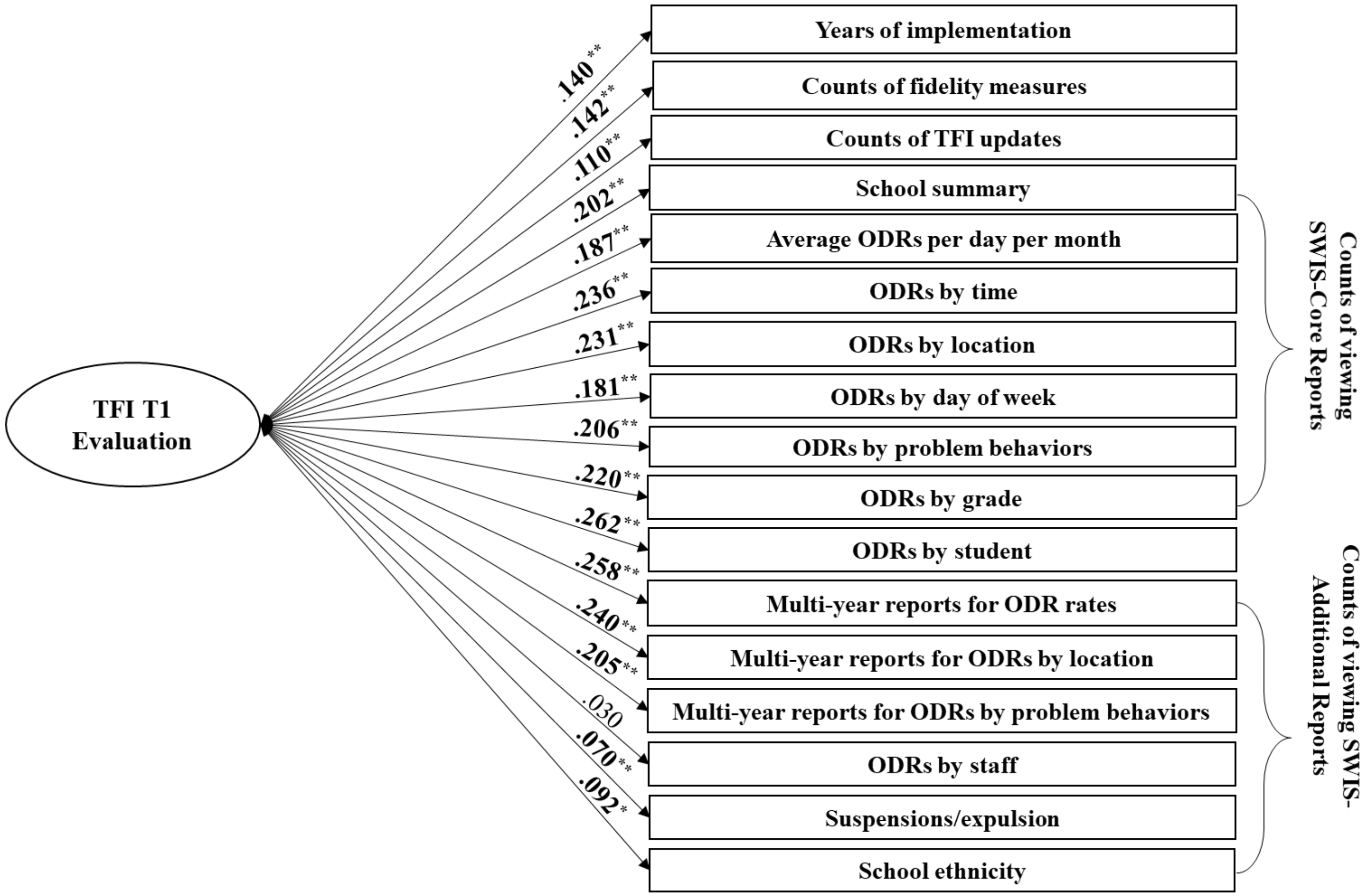

Figure 34. Correlations between the TFI Tier 1 Evaluation subscale and counts of viewing SWIS Report. The covariance paths associated with TFI T1 Evaluation were only depicted. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

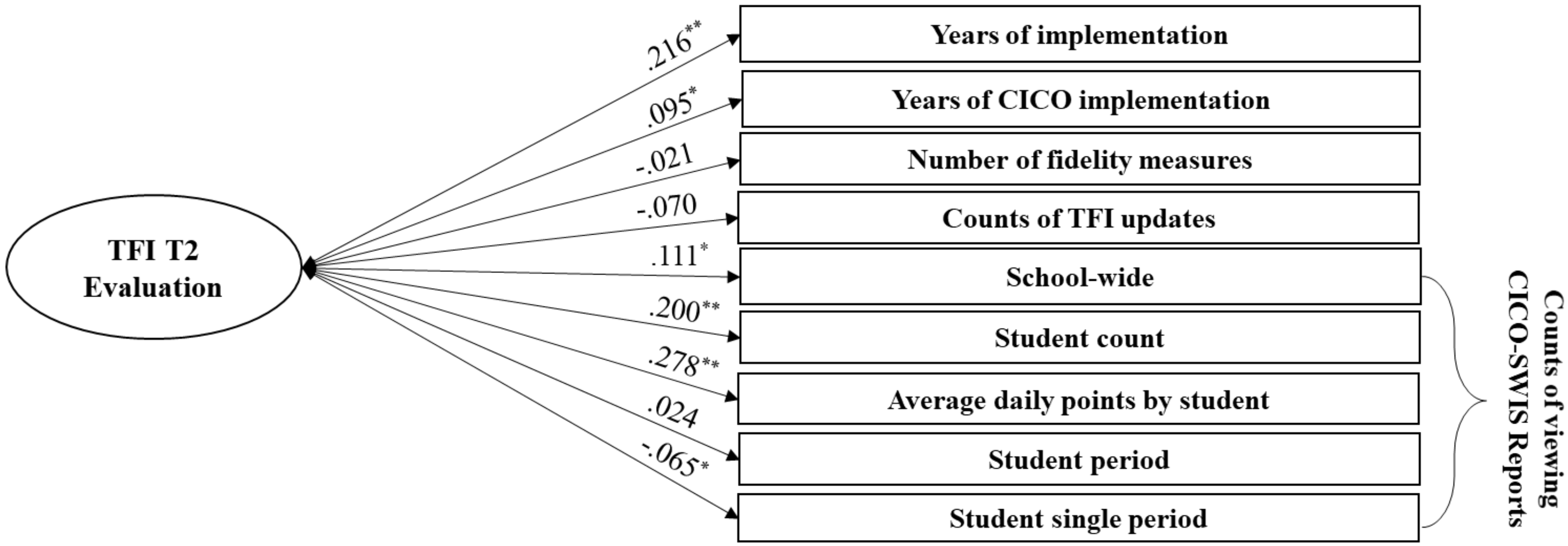

Figure 35. Correlation between the TFI Tier 2 Evaluation subscale and counts of viewing CICO-SWIS Report. The covariance paths associated with TFI T2 Evaluation were only depicted. The statistically significant coefficients ($^{*}p < .05.$, $^{**}p < .01.$) were bolded.

CHAPTER IV

DISCUSSION

The purpose of this study was to examine further the construct validity of TFI score interpretation and usage with an emphasis on predictive and concurrent validity. In particular, the study examined (1) whether the TFI T1 was associated with the square root of major ODRs per 100 students per day, the arcsine of proportions of students with 0 to 1 major ODR, and the racial gap of major ODR rates between African American and non-African American students in the 2016–17 school year; (2) whether TFI T1 was associated with mean-centered proportions of students who met or exceeded standards in state-wide reading and math tests (reading or math achievement) in the 2015–16 school year; (3) whether TFI T2 was associated with the logit of proportions of students enrolled in CICO, the proportion of students who met 80% or more of the goal accomplishments, and the square root of the major ODR rates for students enrolled in CICO in the 2016–17 school year; and (4) whether Evaluation subscales in TFI T1 or TFI T2 were correlated with counts of viewing SWIS or CICO Reports, years of implementing SWIS, years of implementing CICO, counts of TFI administrations, and counts of all fidelity measure administrations during the 2016–17 school year.

Investigating the associations between the TFI and school-level academic and behavioral outcomes provided several findings. Regarding school-wide discipline outcomes, TFI T1 was negatively associated with the square root of major ODRs per 100 students per day and the racial gap of major ODR rates between African and non-African students, and positively with the arcsine of proportions of students with 0 to 1 major ODR in elementary schools, but those associations were not statistically significant except for racial disproportionality. There was a significant and negative association between TFI T1 and the racial gap of major ODR rates between African American and non-African American students in elementary schools. The between-group comparison tests via equality constraint found a moderating effect of grade level on the association of TFI T1 and two behavioral outcome measures involving major ODR rates and the number of students with 0–1 major ODR. In secondary schools, TFI T1 was associated with all behavioral outcomes as hypothesized, but none were statistically significant. In addition, the sensitivity test for the subsample of schools with all tiers of TFI found a significant and negative association between TFI T1 and the square root of major ODR rates in elementary schools, indicating that schools completing all three tiers had lower ODRs when Tier 1 implementation was high.

Regarding academic outcomes, there was a significant and positive association between TFI T1 and reading achievement in elementary schools, a significant moderating effect of years of implementing SWPBIS on the association between TFI T1 and math achievement in elementary schools (stronger fidelity effect on math achievement for elementary schools implementing SWPBIS 7 or more years), and a significant moderating effect of grade level on the association between TFI T1 and reading achievement (stronger effect of fidelity on reading achievement for elementary schools). The sensitivity test for schools assessing all three tiers of TFI found that TFI T1 significantly affected reading- and math-related achievement.

From school-level behavioral outcomes of students enrolled in CICO, TFI T2 scores were negatively associated with % of students with 80% or above goal accomplishment and square root of major ODRs rates for students enrolled in CICO, but none were significant. There was a significant positive association only between TFI T2 scores and logit of proportions of students with CICO daily points. These results were consistently shown in the sensitivity test.

Regarding the Evaluation subscales of TFI T1 and T2, TFI T1 Evaluation was significantly and positively correlated with years of SWPBIS implementation, number of fidelity measure administrations, counts of TFI administrations, and counts of viewing all SWIS Reports except the SWIS ODRs by Staff Additional Report. Also, the TFI T2 Evaluation subscale was significantly positively correlated with years implementing SWPBIS, years implementing CICO, counts of viewing all CICO Reports except for CICO Report – student single period and student period. Specifically, the TFI T2 Evaluation subscale scores were significantly negatively correlated with counts of viewing student single period, but insignificantly positively with counts of viewing student period. Furthermore, TFI T2 was negatively correlated with number of fidelity measure administrations and counts of TFI administrations, but the correlation was not statistically significant.

Associations between TFI T1 and Academic and Behavioral Outcomes

As for Question 1, this study ultimately could not reject the null hypotheses in regard to the associations between TFI T1 and school-wide discipline outcomes—in terms of the rates of major referrals or the proportions of students with 0–1 major ODRs. However, previous studies (Childs et al., 2016; Freeman et al., 2016; Mathews, McIntosh, Frank, & May, 2014; Simonson et al., 2012) have documented the associations between fidelity of SWPBIS implementation and school-wide discipline outcomes. For example, Childs et al. (2016) found significant and negative longitudinal associations between BoQ total scores and the initial level of ODRs, and their results were consistent with those of Flannery et al.(2014), Mathews et al. (2014), and Freeman et al. (2016). As for the inconsistent findings, one explanation may relate to the sampling procedure used in the present study. Because the prior year outcome was tested as the covariate, the subsample of schools with SWIS data for 2 consecutive years was only selected for Question 1, which might account for the unique characteristic of the subsample schools that used SWIS for at least 2 years. As such, autoregressive paths from year to year were involved for all dependent variables, which most strongly explained variance.

Comparing these findings with those from the sensitivity tests provided another explanation. The results revealed that for elementary schools that used all tiers of the TFI, those with a stronger fidelity of SWPBIS implementation at T1 were likely to have significantly lower levels of the square root of major ODR rates. These findings indicated that the associations between fidelity of SWPBIS implementation at T1 and the levels of students' problematic behaviors became stronger for the selected sample of schools that implemented all three tiers than for the larger sample of schools that implemented T1 alone or with T2 or Tier 3. Although the variance of school-discipline measures in such cases cannot be exclusively explained by T1 because of the detected multicollinearity among the tiered fidelity scores, adequate implementation of a universal support system is expected to offer system-level readiness for additional tiers and thus efficient and accurate implementation of the integrated support systems, resulting in improved student behaviors (Lane et al., 2013; Horner & Sugai, 2015). In this context, these results provide initial evidence for theoretical effects of implementing a full MTSS model.

In addition, the results of the present study indicate that a statistically significant association exists between TFI T1 and the risk difference of major ODR rates between African American and non-African American students; this result led to the rejection of the null hypothesis that there were no significant association between fidelity and racial disproportionality. In particular, this result indicates that schools implementing SWPBIS with stronger fidelity are likely to have a reduced level of racial disproportionality in school disciplinary actions. This finding enhances the explanation of Vincent et al. (2011) and the finding of Gage et al.'s (2018) quasi-experimental study: Schools that implement SWPBIS had a significantly smaller disproportionality for African American students in school discipline than for their counterparts. Although no previous studies focused on the relations between the fidelity of SWPBIS implementation and racial disproportionality, one explanation might relate to the use of discipline data (McIntosh, Barnes, Eliason, & Morris, 2014; McIntosh, Ellwood, McCall, & Girvan, 2018), which is an essential component of SWPBIS. Noting that the implicit bias can influence teacher judgment toward school discipline out of conscious awareness, school teams can regularly review the disciplinary data; capture the problem in terms of magnitudes, locations, and/or types of racially unequal referral practices; revise and implement their action plans in a feasible and stepwise approach; and evaluate their implementation (McIntosh et al., 2018). Overall, this study offered a promising finding that supports the importance of implementing core components of PBIS with strong fidelity to improve equity in school discipline.

In regard to Question 2, this study found statistically significant and positive associations between TFI T1 and the mean-centered proportions of students who met or exceeded state-level benchmarks in reading and (more likely past 6 years of SWPBIS implementation) math tests in elementary schools. These results are consistent with prior studies (Gage et al., 2017; Horner et al., 2009) in which schools with BoQ or SET scores equal to or greater than the criterion showed that significantly larger numbers of students met or exceeded state-wide expectations for academic achievement in reading and math. Although there are relatively fewer consistent findings in this regard (e.g., Freeman et al., 2016; Kim et al., 2018), these results confirmed the existing evidence that fidelity relates to academic achievement. Further, the sensitivity test found significant and positive associations between TFI T1 and academic achievement both in reading and math, thereby supporting the function of the fidelity of implementation of universal supports on academic achievement within MTSS.

For Question 1 and Question 2, multigroup modeling was employed (except for the risk difference in Question 1) to specify the parallel models for elementary and secondary schools. However, no significant associations of TFI T1 scores with student outcomes were observed in secondary schools, though many studies (Flannery et al., 2014; Freeman et al., 2016) found negative associations between fidelity and school disciplinary actions. Furthermore, the significant moderating effects of grade levels on the association between fidelity and students' behavioral outcomes were found by the DIFF tests, which compared the regression coefficients with and without equality constraints of student outcomes. In particular, elementary schools showed significantly stronger associations between TFI T1 and the square root of major ODR rates, the arcsine of proportions of students with 0–1 major ODR, and the mean-centered proportions of students who satisfied the state-level benchmark for reading. These results indicate that student performance (either academic or behavioral) might respond to varying degrees of fidelity of implementation more slowly in secondary schools than in elementary schools.

Although such an inconsistency might be related to the relatively small sample size of secondary schools that use the TFI in comparison with elementary schools in this study, another potential explanation concerns the environmental features of secondary schools that might affect the implementation of SWPBIS and their commitment on different student outcomes. As indicated by the results from the measurement models in this study, elementary schools are likely to have significantly higher levels of TFI T1 than secondary schools. Moreover, Flannery, Sugai, and Anderson (2009) argued that secondary schools exhibited complex environments, such as a larger number of staff and students, numerous departments and administrative organizations, and various emphases on academic achievement and postsecondary outcomes, all of which may delay the introduction, implementation, and scale up of SWPBIS.

One assumption of this study was that the associations between fidelity and student outcomes would vary depending on the number of years in which SWPBIS had been implemented. For school-wide disciplinary outcomes, the null hypotheses that there would be no differences in associations between fidelity and student outcomes according to years of implementing SWPBIS could not be rejected, whereas they were rejected for math achievement in elementary schools. Such findings are not consistent with Kim et al.'s (2018) study, indicating the significant associations between fidelity by years of implementing SWPBIS only and Out of School Suspensions per 100 students per days (but not on academic achievement). Those differences might be related to the use of different cutoffs on years of implementing SWPBIS. In particular, Kim et al. (2018)

compared schools that had been implementing SWPBIS for fewer than 3 years with schools that had used it consistently for 3 or more years for academic and behavioral outcomes; however, this study used a dummy variable that indicated whether schools had been using SWPBIS for 7 or more years for academic achievement. In the context of the longitudinal associations between fidelity and academic achievement (Gage et al., 2017), the stronger associations between fidelity and math achievement for schools past 6 years of implementation than those in less than 7 years appeared reasonable.

In addition, the significant interaction terms between fidelity and with other measures were not found in this study, and this result, or lack thereof, requires careful interpretation. As shown in Figure 12 to 15, academic achievement in reading and math appeared higher for schools with stronger fidelity when the schools had been using SWPBIS for 7 or more years, whereas academic achievement was not high when the schools had been using SWPBIS for fewer than 7 years. These findings suggest that SWPBIS might need to be used for many years to promote stronger association between implementation fidelity and reading achievement because of the nature of the subject (Kim et al., 2018). Similarly, the insignificant moderation of association between TFI T1 and school discipline by years implementing might be explained by the use of the certain cutoff point (year 1–4/5 or more) on the years of using SWPBIS. For example, using multiple cutoffs for more than two experience groups would be applicable, as indicated by recent research (Gage et al., 2018) in which three experience groups (1-2/3-5/6-10 years of implementing SWPBIS) were compared with the propensity of score-matched comparison schools (without any SWPBIS training). Gage et al. (2018) found that schools with fidelity in 3–5 years of implementation showed significantly lower OSSs and proportions of students with either only one or more than one OSSs than the reference group, whereas schools with fidelity in 1–2 years or 6–10 years had a significantly smaller number of students with only one OSS. Considering that those differences might be related to the varying levels of fidelity (between or within experience groups) and the years of implementation experiences, various cutoffs can be used to determine whether the association between fidelity and student outcomes is moderated by years of implementation.

In addition, years of SWPBIS implementation were tested as the sole covariate in this study, and it was found that (if the effects of prior year outcomes and fidelity were controlled for) the schools that had been using SWPBIS for fewer than 5 years were likely to have higher ODR rates almost equally in elementary and secondary schools than those that had been using SWPBIS for 5 or more years. As indicated by prior studies (Childs et al., 2016; Gage et al., 2018), such a result suggests that schools implementing an SWPBIS program for at least 3 – 5 years are likely to have reduced levels of school discipline. In addition, for all measures in Question 1 and 2, multiple demographic covariates were controlled. As a result, urbanicity was significantly negatively associated with major ODR rates and positively associated with the proportions of students with 0 – 1 major ODRs for elementary and secondary schools. The percentage of students who were eligible for free or reduced-price lunch was significantly negatively associated with the proportion of students with 0 – 1 major ODRs and reading and math achievement, for both elementary and secondary schools. In addition, middle schools that had Grades 6 – 8 were significantly associated with risk difference in a positive direction but only for secondary schools.

Associations between TFI T2 and Behavioral Outcomes

Regarding Research Question 3, there was only a significant and positive association of TFI T2 on the proportions of students with CICO daily points. Such results were supported by Hawken et al.'s (2015) descriptive study, in which schools that scored 70% or higher in the Foundations and Targeted ISSET showed larger proportions of students with CICO than those with lower proportions of the criterion. Schools that built a strong fidelity of implementation at SWPBIS T2 could facilitate the implementation procedure of CICO in an efficient, effective, and feasible manner for the CICO coordinators and school teams to promptly arrange new students in need of the CICO routine, rearrange the existing students enrolled in CICO to the matching interventions with varying intensity, or discontinue the CICO if appropriate. Similarly, the present study found that school size was a predictor of the proportion of students with CICO daily points, as indicated by Hawken et al. (2015). Schools with higher proportions of enrolled students might enroll smaller proportions of students in CICO because of the limitations in the proportion of available students whom one CICO coordinator could afford to support. In this context, Hawken et al. (2015) recommended that schools with higher numbers of enrolled students should identify enough coordinators and facilitators (as well as internal or external personnel) to effectively serve approximately 7% – 15 % of students in CICO. Overall, these results indicate that schools that apply T2 with a strong fidelity are likely to have larger proportions of students in CICO.

However, this study found that TFI T2 was not significantly related to the proportion of students who met 80% or more of daily point goals or the major ODR rates for students enrolled in CICO. Although there is strong evidence that CICO effectively reduces problematic behaviors and increases academic engagement (Mitchell, Adamson, & McKenna, 2017), TFI T2 defines the core elements of targeted support more than CICO does; thus, the variances in the two direct outcome measures from students enrolled in CICO might not be significantly explained by TFI T2. Nevertheless, such results require careful interpretation. In the context of the characteristics of the data sample used in this study, it should be noted that most schools (approximately 52.98%) had been using CICO for 4 or fewer years, although schools in year 1 were excluded. During the initial period in which newly introduced innovations settle down to adequate fidelity through repeated trials and errors, it is difficult to ascertain a clear pattern of association between the fidelity of implementation and student outcomes (Nese et al., 2018).

In addition, the interaction terms between fidelity and the years of CICO implementation (indicated by whether schools had been using it for 2–3 years)were not statistically significant in this study. However, the years of CICO implementation of the sample ranged from 2 to 8, which appeared to be a smaller range than that of schools that implemented T1, according to Question 1 and 2. Using different cutoffs (e.g., shorter or longer years) with larger samples with varying years of implementation might have produced the unique variations in outcome measures and the associations between fidelity and those measures. Regarding this point, many review studies (Hawken et al., 2014; Maggin, Zurheide, Pickett, & Baillie, 2015; Mitchell et al., 2017) noted a relatively small number of group-based studies in comparison to single subject studies. In particular, only a few large-scale studies have examined fidelity of implementation as one predictor of student outcomes, which narrows further discussion of the current evidence on fidelity effects on CICO. Regarding demographic covariates, the proportions of students eligible for free or reduced-priced lunch and elementary schools were negatively associated with the percentage of students who met 80% or more of the goal accomplishments.

Correlations between Evaluation Subscale and Relevant Variables

This study found that the TFI T1 Evaluation subscale was positively and weakly (but significantly) correlated with years implementing SWPBIS, the number of TFI administrations, the number of fidelity measures completed, and all numbers of viewing SWIS reports. Consistent with previous studies (Nese et al., 2018; Schaper, McIntosh, & Hoselton, 2016), schools with more experiences of implementing SWPBIS tended to report higher scores in the TFI T1 Evaluation subscale. Also, schools with higher scores in the TFI T1 Evaluation subscale tended to report that they used more types of fidelity measures and completed the TFI more frequently, which are matched to two items (14 and 15) in the TFI T1 Evaluation subscale. Those items assessed whether school teams used fidelity measures at least annually and whether they evaluated and reported to the stakeholders the effectiveness of SWPBIS via documenting fidelity and student outcome data (Algozzine et al., 2014). Regarding Items 12 and 13, schools with higher scores in TFI T1 Evaluation tended to review SWIS reports more frequently. These two items assessed whether school teams accessed the discipline data report systems and whether they reviewed and used those summarized data for decision-making (Algozzine et al., 2014). However, TFI T1 Evaluation was positively but insignificantly correlated with counts of viewing the SWIS ODRs by Staff Additional Report. One explanation might be that this filter could be accessed only by administrators or SWIS coordinators due to possible negative impacts on peer reputation although this tool provides useful information regarding staff's referral practices (Eliason & Morris, 2015). Overall, these results indicate that the TFI T1 Evaluation subscale measures what it is intended to measure.

Similarly, the TFI T2 Evaluation subscale was positively significantly associated with years of SWPBIS and CICO implementation as hypothesized. As it is typically recommended that schools introduce CICO at least 2–3 years after meeting and maintaining the criterion of fidelity at T1 (Hawken et al., 2015), schools with sustained fidelity of SWPBIS implementation are likely to have stronger foundation system for effective and efficient CICO implementation, leading to higher scores in TFI T2 Evaluation. In addition, the TFI T2 Evaluation subscale was positively associated with counts of viewing most CICO–SWIS Reports.

Contrary to expectations, the TFI T2 Evaluation subscale was not significantly correlated with viewing the Individual Student Period report (i.e., the average percentage of points gained by individual students each period within a specified range of dates) and was significantly but negatively correlated with counts of viewing the Individual Student Single Period (i.e., the average percentage of points earned by individual students during a particular period each day). Compared with other daily, individual-level data—such as Average Daily Points by Student and Individual Student Count (i.e., the percentage of points per day by individual students)—those types of CICO–SWIS Reports are less frequently generated, as indicated by Table 28. The uses of those measures might depend on an individual student's need for monitoring with a higher intensity and his or her progress, which might be explained by CICO fidelity rather than the T2 support system.

Furthermore, if schools with strong fidelity have more students enrolled in CICO, the coordinators might have less time to consider the data in detail, leading them to focus more on students' daily points. Although school-wide data might be monitored less frequently than individual data, school teams might need to focus on the school-wide patterns of outcomes of students enrolled in CICO; by doing so, the positive correlations might appear reasonable.

Implications for Practitioners and Professionals

First, this study enhanced the evidence for the construct validity of TFI score interpretations for their purposed uses, such as self-monitoring or annual evaluation (Algozzine et al., 2014), by linking the TFI to the equivalent measures and relevant outcome data. In particular, the study empirically justified that (1) elementary schools with higher TFI scores at T1 are likely to show not only lower rates of school discipline if all three tiers are used but also higher levels of academic achievement; that (2) those elementary schools are also likely to have reduced levels of risk difference in major ODR rates between African American and non-African American students; that (3) schools with higher TFI scores at T2 were likely to have more students receiving CICO interventions; and that (4) schools with higher TFI T1 and T2 Evaluation subscale scores tend to view data, such as student outcome data and fidelity data, more often. Thus, they can provide school teams with useful information about not only the construct validity but also the utility of the TFI to help their selections and uses of the TFI fitting to their needs, across various implementation stages.

In addition, this study highlighted the long-term associations between SWPBIS fidelity at T1 and academic achievement. As indicated by many studies (McIntosh, Flannery, Sugai, Braun, & Cochrane, 2008; McIntosh, Horner, Chard, Dickey, & Braun, 2008; Nelson, Benner, Lane, & Smith, 2004), reducing problematic behaviors in classrooms can increase academic engagement and thereby improve academic achievement, whereas deficiencies in academic skills can cause students to engage in problematic behaviors. Given the close relationship between behavior and learning, researchers have emphasized that academic and behavioral support should be integrated into the response to intervention (RTI) system because of the shared elements of SWPBIS with academic RTI, such as team-driven, data-based decision-making or the use of evidence-based practices (Bohanon, Goodman, & McIntosh, 2011; Hawken, Vincent, & Schumann, 2008). Although the TFI focuses only on behavioral support systems, the positive association between TFI T1 fidelity and academic achievement implies that the combined academic and behavioral interventions can provide students better gains than implementing the parts separately.

Given the comprehensive scopes and depths of behavioral supports in MTSS, multiple years of effort are required to change and maintain a school and its classroom practices enough to produce the desired effects on student outcomes, as indicated by previous studies (Bradshaw et al., 2012; Gage et al., 2016; Kim et al., 2018); this need for time and effort implies that patience and consistency are required when implementing SWPBIS. For the sustained use of SWPBIS at the school level, Bohanon et al. (2011) recommended that new initiatives be connected to the existing priorities at the school or district levels by developing relevant practices (e.g., integrating academic RTI and SWPBIS) in parallel. In addition, McIntosh, Mercer, et al. (2016) noted that the capacities of schools should be systematically developed by the upper levels of authority such as states or districts so that various support systems such as policy, technical supports, evaluation, and funding can be implemented in an interconnected way. In particular, intensive technical supports may be needed during the initial period to increase fidelity of implementation to criteria and to motivate and sustain a school staff's commitment to the school-wide initiative (McIntosh, Mercer, et al., 2016; Nese et al., 2018).

In addition, the negative and significant associations between TFI T1 and major ODR rates of the elementary schools with all TFI T1, T2, and T3 scores supported the integrated function of three-tiered support systems, which was also shown in both reading and math achievement. In past studies (Horner & Sugai, 2015; McIntosh & Goodman, 2016), emphasis was placed on implementing universal prevention efforts with sustainable fidelity to prepare for scaling to the advanced tiers, because high-quality prevention efforts enable students to meet a positive, predictable, and consistent environment before the students display the challenging problems; thus, the positive effects of SWPBIS on student outcomes are maximized. In particular, the consistency between classroom practice and the school-wide systems is a key part of the pathways that link the integrated systems of academic and behavioral supports to students' outcomes by changing teachers' behavioral management and instructions in classrooms (Darch, Kame'enui, & Crichlow, 2003; McIntosh & Goodman, 2016). Similarly, Bradshaw et al. (2012) found that schools implementing both T1 and T2 behavioral supports with training and coaching together were more likely to provide students with classroom-based behavioral support than schools with only traditional training services, thereby indicating that consistent and hands-on external support is needed not only for school teams but also for teachers' skill development.

However, evidence for how TFI T1 is related to academic and behavioral outcomes was found only in elementary schools in this study. Because of comparatively low TFI scores in secondary schools, there is a strong need not only for professional development opportunities for adherence to core features of SWPBIS but also for the consideration of their own contextual characteristics of high schools (Flannery & Kato, 2017). In particular, Flannery and Kato (2017) proposed the use of multiple adaptation strategies, such as the following: building a leadership team, representing content-driven or service-based organizations of high schools, promoting active participation of adolescent students, using comprehensive and team-based efficient communication systems for high numbers of students and staff, and considering attendance and academic data in the context of strongly emphasizing academic achievement and preventing students from dropping out.

In addition, this study noted the role of fidelity of SWPBIS in enhancing not only school discipline but also equity in school discipline. To intervene with school practices and policies to exacerbate the effects of explicit and implicit biases, McIntosh, Girvan, Horner, and Smolkowski (2014) proposed multiple strategies based on the effective implementation of SWPBIS. In particular, they emphasized the active use of school discipline data to moderate the effects of either explicit or implicit biases in recognition that the ongoing collection, review, and report of the racially disaggregated data can enhance the commitment of school personnel and enables school teams to identify the situations in which disproportionate discipline is likely to occur, to clarify the relevant disciplinary procedures (e.g., definitions of subjective ODRs), and to train school teachers to follow a culturally responsive practice. Despite this importance, McIntosh, Eliason, Horner, and May (2014) reported that schools had not viewed the SWIS School Ethnicity Report as much as other SWIS Report tools, which requires ongoing professional development with emphasis on analyzing the disaggregated discipline data in various aspects and using data-viewing tools (e.g., SWIS Drill Down). Also, McIntosh et al. (2014) argued that different levels of authorities from school to state levels need to document the schools' accountability for equity in school practice via a) an explicit mission statement that the school is committed to disciplinary equity, b) a clearly defined and feasible procedure (e.g., professional development, data systems, hiring) to address any disproportionality, and c) the evaluation of systems to ensure accountability.

Furthermore, the positive correlations between evaluation scores of either the TFI T1 or T2 and how frequently school teams reviewed student data also highlight the importance of establishing a data system as a pivotal facilitator of SWPBIS implementation to promote ongoing data collection and reviews for decision-making—particularly in recognition of the empirically validated effectiveness of fidelity of SWPBIS on student outcomes (Childs et al., 2015; Freeman et al., 2015). In particular, several researchers (e.g., Kennedy et al., 2009; McIntosh et al., 2013) recommended that an efficient, feasible, and compatible database system be used to help school teams promptly access, summarize, and visualize student outcome data at the school or individual levels. In addition, data sharing with different stakeholders (e.g., school staff and district or state administrators) across various levels of authority should improve their sense of urgency for students' behaviors and learning, as well as priorities on the innovation, thereby encouraging and sustaining their commitment, which is critical for long-term implementation and success (Bohanon et al., 2011).

Regarding TFI T2, this study noted that to improve the fidelity of T2 implementation, schools need strong capacity-building support. To consistently and promptly serve an adequate proportion of students in CICO, it is necessary that T2 systems be used with fidelity to the criterion over time, which requires that districts or states or both provide schools sufficient resources, such as personnel, funding, and technical support. In this regard, previous large-sized studies on CICO interventions placed an emphasis on professional development (Hawken et al., 2015; Bradshaw et al., 2012) for school teams and coaches at T1 or T2 or both. In particular, retraining T1 systems commonly precedes T2 CICO training, in the sense that the routine of CICO is extended from school-wide expectations by adding frequent and regular monitoring and feedback.

Another aspect of school capacity for efficient and effective implementation of CICO pertains to having an adequate number of coordinators within schools. These coordinators must be part of the school staff, appropriately trained, and continuously coached to facilitate the CICO routine for 7%–15% of students with increased intensity (Hawken et al., 2015). Their accurate implementation of CICO requires the T2 coaches to ensure that the teachers correctly understand CICO intervention (as a proactive prompt to prevent the occurrence of problematic behaviors), particularly in the context of functional perspectives (Hawken et al., 2015; McIntosh & Goodman, 2016). As CICO facilitators, teachers' knowledge and skills in behavioral management are considered essential to integrating T2 into school-wide support systems (McIntosh & Goodman, 2016). In addition, considering the functions of behaviors within their surrounding contexts can guide school teams' and teachers' decision-making processes for identifying and referring students who need additional support, selecting appropriate interventions, and monitoring or evaluating them. In this context, Bradshaw et al. (2012) evaluated the effects of T2 coaching, revealing from the teacher-level repeated measure analyses that teachers from schools using T1 and T2 with ongoing T2 coaching reported significantly large decreases in the proportions of students with academic difficulties and students receiving special education services than did teachers from schools without ongoing T2 coaching.

Limitations and Recommendations for Future Research

This study had several limitations. First, this cross-sectional study did not have a clear order of precedence between the independent and dependent variables, although the prior year outcomes and years of implementing were controlled for. This predictive limitation requires that the cause–effect relationships between fidelity and student outcomes be interpreted carefully. Thus, future research should consider different longitudinal research designs and analytic approaches. Longitudinal research (e.g., cross-lagged panel model, parallel latent growth model) allows researchers to examine sequences of observations associated with fidelity and student outcomes over multiple years, which can suggest the cause-effect relationship between two variables. In addition, the unit of analysis in this study was the school, which might have caused the relatively small variations in the aggregated student outcomes at the school level. Given the clustered structure of this study's student outcome data, future research should use multilevel models that involve the student and school levels, as well as those above.

In regard to the samples, the range of TFI scores was restricted at the upper bound, which may have affected the results of correlations or regressions in this study. In specific, the majority of schools met or exceeded 70% at T1, or the criterion score of the TFI, and co-variances between fidelity (with a restricted range) and student outcomes, cannot represent a broad pattern on a wide range of fidelity scores. Therefore, the findings of this study require careful interpretation. In addition, the majority of the schools in the sample were regular elementary schools. Because only a few middle and high schools were included, this study could not ultimately provide evidence of the association between fidelity and secondary student outcomes. Given that complexity in analytic models is unavoidable because schools have various characteristics (e.g., years of implementation and demographic characteristics), sample sizes must be large enough to reduce bias, in the context of diverse population groups and school types for additional validation studies.

In addition, more validation research needs to be conducted with various outcome data about each tier of SWPBIS. This study did not evaluate the associations between TFI T1 or TFI T2 and serious problematic behaviors of students with high rates of major ODRs or suspensions. As indicated by previous research (Childs et al., 2016; Simonson et al., 2012), those measures (e.g., total events of OSS or ISS and the percentage of students with suspensions) might be more responsive to varying scores of the TFI. In addition, this study used risk difference (a commonly used comparative metric of racial disproportionality), or the absolute size of the difference between groups in the risk index. Although risk difference does not inform the relative magnitude of ODR rates between target and reference groups as risk ratio does, it is applicable to schools with zero major ODR rates for the reference group (which is the denominator). Despite the relative stability and the wider usability of the risk difference, Girvan et al. (2018) noted that no single metric can fully explain racial disproportionality in school discipline. As such, future research with various measures is needed to expand the interpretation of the predictive validity on disciplinary equity and to compensate emerging analytic limitations. In addition, because of the low stability of schools with very few students in student subgroups (Girvan et al., 2018), only schools with 10 or more students in both groups were enrolled for analyses; thus, the fidelity effects on disproportionality must be interpreted cautiously, which requires the applications of various research methods (e.g., qualitative or single subject research). Regarding the concurrent validity of the TFI T2 Evaluation subscale, this study did not use the fidelity data of CICO intervention. Thus, future research is needed to test whether TFI T2 Evaluation subscale scores are associated with whether school teams collected and how frequently they measured the fidelity of CICO intervention, as well as the extent of actual fidelity.

Regarding the measurement models, this study used subscale indicators to secure the analytic power because of the model complexity. To more accurately estimate the factor variances from the full sets of item parcels, future research should specify item-level measurement models if a sufficient size of the sample is secured. In specific, three-factor item-level measurement models can be specified aligned to the theoretical framework of the TFI. In addition, this study evaluated each single scale separately because of the multicollinearity among the three tier scales. However, dropping other scales did not remove the shared variations among the three scales, so careful interpretation is again warranted. Although the nature of MTSS restricted the estimation of the unique variances explained by each tier, one way to address this issue might be to use an alternative estimator, such as partial linear squares (PLS-SEM), to reduce bias via maximizing the variance accounted for in dependent variables, rather than explain between-indicator covariation (Hair et al., 2011). Another way might be to use a multidimensional model that estimates one factor of the TFI as the whole construct, while keeping the nature of MTTS. This study did not examine TFI T3 because accessibility to relevant outcome data was limited (e.g., I-SWIS). Thus, future research is needed to determine the predictive and concurrent validity of TFI T3 with sufficiently sized samples.

Conclusion

This study expanded the empirical evidence to explain the construct validity for the TFI score interpretations and uses by linking TFI T1 or T2 scores to relevant external measures. Particularly, the findings of this study indicate that schools that use SWPBIS and have higher scores at TFI T1 are likely to achieve better student outcomes, either in academic achievement or in school discipline, and that those with higher scores for TFI T2 are likely to serve a larger proportion of students in CICO. Moreover, the result of the sensitivity tests with the subsample of schools that used all tiers of the TFI showed the effects of TFI T1 within multi-tiered support on academic achievement and school disciplines. Adding to the previous validation studies, this study concluded that the TFI (particularly TFI T1 and T2) is highly expected to measure what they are intended to measure for their goals. Given that there is no complete status of construct validity for any interpretation and use of the test scores, ongoing validation efforts need to continue with various hypothesized arguments and relevant research designs.

REFERENCES CITED

Algozzine, B., Horner, R. H., Sugai, G., Barrett, S., Dickey, S. R., Eber, L., Kincaid, D., et al. (2010). *Evaluation blueprint for school-wide positive behavior support*. Eugene, OR: National Technical Assistance Center on Positive Behavior Interventions and Support. Retrieved from www.pbis.org.

Algozzine, B., Barrett, S. Eber, L., George, H., Horner, R., Lewis, T., Putnam, B., Swain-Bradway, J., McIntosh, K., & Sugai, G (2014). *School-wide PBIS Tiered Fidelity Inventory*. Eugene, OR: OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. Available at http://www.pbisapps.org.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anderson, C. M., & Borgmeier, C. (2010). Tier II interventions within the framework of school-wide positive behavior support: Essential features for design, implementation, and maintenance. *Behavior Analysis in Practice*, *3*(1), 33-45.

Anderson, C.M., Lewis-Palmer, T., Todd, A.W., Horner, R.H., Sugai, G., and Sampson, N.K (2012). *Individual Student Systems Evaluation Tool (ISSET v. 3.0)*. Eugene, OR: Educational and Community Supports. Available at http://www.pbisapps.org.

Asparouhov, T. (2009, January 22). Re: Offset in poisson regression. Retrieved from http://www.statmodel.com/discussion/messages/23/781.html

Asparouhov, T. & Muthén, B. (2010, August). *Weighted least squares estimation with missing data*. (Technical Report). Retrieved from http://ww.statmodel2.com/download/GstrucMissingRevision.pdf

Barrett, S. B., Bradshaw, C. P., & Lewis-Palmer, T. (2008). Maryland statewide PBIS initiative systems, evaluation, and next steps. *Journal of Positive Behavior Interventions*, *10*(2), 105-114.

Bassett, S. S., Stein, L. A. R., Rossi, J. S., & Martin, R. A. (2016). Evaluating measures of fidelity for substance abuse group treatment with incarcerated adolescents. *Journal of substance abuse treatment*, *66*, 9-15.

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D. & Czajkowski, S. (2004). Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*, *23*(5), 443.

Bloomfield, A. (2015). *An investigation of the content and concurrent validity of the School-wide Evaluation Tool.* (Unpublished doctoral dissertation). Temple University, Philadelphia.

Bohanon, H., Goodman, S., & McIntosh, K. (2011). Integrating academic and behavior supports within an RtI framework, Part 1: General overview. Retrieved June 30, 2014 from http://www.rtinetwork.org/learn/behavior-supports/integrating-behavior-andacademic-supports-general-overview

Boneshefski, M. J., & Runge, T. J. (2014). Addressing disproportionate discipline practices within a school-wide positive behavioral interventions and supports framework: a practical guide for calculating and using disproportionality rates. *Journal of Positive Behavior Interventions*, *16*(3), 149-158.

Bradshaw, C. P., Debnam, K., Koth, C. W., & Leaf, P. (2009). Preliminary validation of the implementation phases inventory for assessing fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions*, *11*(3), 145-160.

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*(3), 133-148.

Bradshaw, C. P., Pas, E. T., Goldweber, A., Rosenberg, M. S., & Leaf, P. J. (2012). Integrating school-wide positive behavioral interventions and supports with tier 2 coaching to student support teams: The PBIS plus model. *Advances in School Mental Health Promotion*, *5*(3), 177-193.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, **2**(3), 171–174.

Brown, T. (2006). *Confirmatory factor analysis for applied research.* New York: Guildford.

Burns, G. L., Walsh, J. A., Gomez, R., & Hafetz, N. (2006). Measurement and structural invariance of parent ratings of ADHD and ODD symptoms across gender for American and Malaysian children. Psychological Assessment, 18(4), 452.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 40, doi:10.1186/1748-5908-2-40.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199-218.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468-508.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. Structural Equation Modeling, 12, 471– 492.

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.

Cizek, G. J. (2012). Defining and distinguishing validity: interpretations of score meaning and justification of test use. *Psychological methods, 17*(1), 31-43.

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.

Crone, D. A., Hawken, L. S., & Horner, R. H. (2010). Responding to problem behavior in schools: The Behavior Education Program (2nd ed.). New York, NY: Guilford Press.

Cubilo, J. (2014). Argument-Based Validity in Classroom and Program Contexts: Applications and Considerations. *Shikent Research Bulletin*, 18-24.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical psychology review*, *18*(1), 23-45.

Darch, C. B., Kame'enui, E. J., & Crichlow, J. M. (2003). *Instructional classroom management: A proactive approach to behavior management* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

DeLuca, C. (2011). Interpretive validity theory: mapping a methodology for validating educational assessments, *Educational Research, 53*(3), 303-320.

Determining Significant Disproportionality, 34 C.F.R. § 300.647 (2016).

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health education research*, *18*(2), 237-256.

Eber, L., Breen, K., Rose, J., Unizycki, R. M., & London, T. H. (2008). Wraparound as a tertiary level intervention for students with emotional/behavioral needs. *Teaching Exceptional Children*, *40*(6), 16-22.

Fairbanks, S., Simonsen, B., & Sugai, G. (2008). Classwide secondary and tertiary tier practices and systems. *Teaching Exceptional Children*, *40*(6), 44-52.

Fairbanks, S., Sugai, G., Guardino, D., & Lathrop, M. (2007). Response to intervention: An evaluation of a classroom system of behavior support for second grade students. *Exceptional Children*, *73*, 288-310.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on social work practice*, *19*(5), 531-540.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation research: A synthesis of the literature (Louis de la Parte Florida Mental Health Institute Publication No. 231). Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

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.

Flannery, K. B., & McGrath Kato, M. (2017). Implementation of SWPBIS in high school: Why is it different? *Preventing School Failure: Alternative Education for Children and Youth*, *61*(1), 69-79.

Flannery, K. B., Sugai, G., & Anderson, C. M. (2009). School-Wide Positive Behavior Support in High School. Journal of Positive Behavior Interventions, 11(3), 177–185.

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2006). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior therapy*, *36*(1), 3-13.

Freeman, J., Simonsen, B., McCoach, D. B., Sugai, G., Lombardi, A., & Horner, R. (2016). Relationship Between School-Wide Positive Behavior Interventions and Supports and Academic, Attendance, and Behavior Outcomes in High Schools. *Journal of Positive Behavior Interventions*, *18*(1), 41-51.

Gage, N. A., Grasley-Boy, N., Peshak George, H., Childs, K., & Kincaid, D. (2018). 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*, 1098300718768208.

Gage, N. A., Leite, W., Childs, K., & Kincaid, D. (2017). Average treatment effect of school-wide positive behavioral interventions and supports on school-level academic achievement in Florida. *Journal of Positive Behavior Interventions*, *19*(3), 158-167.

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders*, 37(2), 55-77.

Girvan, E. J., McIntosh, K., & Smolkowski, K. (2018). Tail, Tusk, and Trunk: An Examination of What Different Metrics Reveal about Racial Disproportionality in School Discipline.

Gresham, F. M. (2009). Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review*, *38*(4), 533-541.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, *19*(2), 139-152.

Hawken, L. S., Bundock, K., Barrett, C. A., Eber, L., Breen, K., & Phillips, D. (2015). Large-Scale Implementation of Check-In, Check-Out A Descriptive Study. *Canadian Journal of School Psychology, 30*(4), 304 – 319.

Hawken, L. S., Bundock, K., Kladis, K., O'Keeffe, B., & Barrett, C. A. (2014). Systematic review of the check-in, check-out intervention for students at risk for emotional and behavioral disorders. *Education and Treatment of Children, 37*(4), 635-658.

Hawken, L. S., Vincent, C. G., & Schumann, J. (2008). Response to intervention for social behavior. *Journal of Emotional and Behavioral Disorders, 16*, 213–225.

Hooper, D., Coughlan, J., Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. Electronic Journal of Business Research Methods, 6(1), 53-60.

Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. E. (1987). Taking charge of change. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Horner, R. H., Sampson, N. K., Anderson, C. M., Todd, A. W., & Eliason, B. M. (2013). *Monitoring advanced tiers tool*. Eugene, OR: Educational and Community Supports. Available at http://www.pbisapps.org.

Horner, R. H., & Sugai, G. (2015). School-wide PBIS: an example of applied behavior analysis implemented at a scale of social importance. *Behavior Analysis in Practice, 8*(1), 80-85.

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

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

Horner, R. H., 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.

Irvin, L. K., Horner, R. H., Ingram, K., Todd, A. W., Sugai, G., Sampson, N. K., & Boland, J. B. (2006). Using office discipline referral data for decision making about student behavior in elementary and middle schools: an empirical evaluation of validity. *Journal of Positive Behavior Interventions*, *8*(1), 10-23.

Irvin, L. K., Tobin, T. J., Sprague, J. R., Sugai, G., & Vincent, C. G. (2004). Validity of office discipline referral measures as indices of school-wide behavioral status and effects of school-wide behavioral interventions. *Journal of Positive Behavior Interventions, 6*(3), 131-147.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin, 112*(3), 527-535.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp17-64). Westport, CT: Praeger.

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448.

Kazdin, A. E. (1986). Comparative outcome studies of psychotherapy: Methodological issues and strategies. *Journal of Consulting and Clinical Psychology, 54*, 95–105.

Kennedy, M. J., Horner, R. H., McNelly, D., Mimmack, J., Sobel, D., Tillman, D. R., (2009). Data-based decision making in high schools: Informed implmentation of school-wide positive behavior support. In B. Flannery & G. Sugai (Eds.), *SWPBS implementation in high schools: Current practice and future directions.* (pp. 81-114). University of Oregon.

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.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher*, *34*(8), 3-13.

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

Kline, R.B. (2005), Principles and Practice of Structural Equation Modeling (2nd Edition ed.). New York: The Guilford Press.

Lane, K. L., Oakes, W. P., Menzies, H. M., Oyer, J., & Jenkins, A. (2013). Working within the context of three-tiered models of prevention: Using schoolwide data to identify high school students for targeted supports. *Journal of Applied School Psychology*, *29*(2), 203-229.

Lane, K. L., Rogers, L. A., Parks, R. J., Weisenbach, J. L., Mau, A. C., Merwin, M. T., & Bergman, W. A. (2007). Function-based interventions for students who are nonresponsive to primary and secondary prevention efforts illustrations at the elementary and middle school levels. *Journal of Emotional and Behavioral Disorders*, *15*(3), 169-183.

Lewis-Palmer, T., Todd, A. W., Horner, R. H., Sugai, G., & Sampson, N. K. (2003). *Individual Student Systems Evaluation Tool (ISSET)*. Eugene, OR: Educational and Community Supports. Available at http://www.pbisapps.org.

Maggin, D. M., Zurheide, J., Pickett, K. C., & Baillie, S. J. (2015). A systematic evidence review of the Check-In/Check-Out program for reducing student challenging behaviors. *Journal of Positive Behavior Interventions*, 17(4), 197 – 208.

Maslowsky, J., Jager, J., & Hemken, D. (2015). Estimating and interpreting latent variable interactions: A tutorial for applying the latent moderated structural equations method. *International Journal of Behavioral Development*, *39*(1), 87-96.

Massar, M., McIntosh, K., Mercer, S. H., & Hoselton, R. (in press). Factor analysis of a fidelity of implementation measure for social behavior systems. *Remedial and Special Education.*

Mathews, S., McIntosh, K., Frank, J. L., & May, S. L. (2014). Critical features predicting sustained implementation of school-wide positive behavioral interventions and supports. *Journal of Positive Behavior Interventions*, *16*(3), 168-178.

May, S.L., Ard, W. I., Todd, A.W., Horner, R.H., Glasgow, A., Sugai, G., & Sprague, J.R. (2006). School-Wide Information System. Educational and Community Supports, University of Oregon, Eugene, OR.

McIntosh, K., Barnes, A., Eliason, B., & Morris, K. (2014). *Using discipline data within SWPBIS to identify and address disproportionality: A guide for school teams.* OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. www.pbis.org.

McIntosh, K., Filter, K. J., Bennett, J. L., Ryan, C., & Sugai, G. (2010). Principles of sustainable prevention: Designing scale-up of School-wide Positive Behavior Support to promote durable systems. *Psychology in the Schools*, *47*(1), 5-21.

McIntosh, K., Flannery, K. B., 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.

McIntosh, K., Horner, R. H., Chard, D. J., Dickey, C. R., & Braun, D. H. (2008). Reading skills and function of problem behavior in typical school settings. *The Journal of Special Education*, *42*(3), 131-147.

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. R. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions*, *11*(2), 82-93.

McIntosh, K., Campbell, A. L., & Carter, D. R. Zumbo, BD (2009). Concurrent validity of office discipline referrals and cut points used in schoolwide positive behavior support. *Behavioral Disorders*, *34*(2), 100-113.

McIntosh, K., Eliason, B. M., Horner, R. H., & May, S. L. (2014). Have schools increased their use of the SWIS School Ethnicity Report? PBIS evaluation brief. Eugene, OR: Educational and Community Supports.

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.

McIntosh, K., Girvan, E. J., Horner, R., & Smolkowski, K. (2014). Education not incarceration: A conceptual model for reducing racial and ethnic disproportionality in school discipline.

McIntosh, K., & Goodman, S. (2016). *Integrated multi-tiered systems of support: Blending RTI and PBIS*. Guilford Publications.

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.

McIntosh, K., Mercer, S. H., Hume, A. E., Frank, J. L., Turri, M. G., & Mathews, S. (2013). Factors related to sustained implementation of schoolwide positive behavior support. *Exceptional Children*, *79*(3), 293.

McIntosh, K., Mercer, S. H., Nese, R. N., Strickland-Cohen, M. K., & Hoselton, R. (2016). Predictors of sustained implementation of school-wide positive behavioral interventions and supports. *Journal of Positive Behavior Interventions*, *18*(4), 209-218.

McIntosh, K., Mercer, S. H., Nese, R. N., & Ghemraoui, A. (2016). Identifying and Predicting Distinct Patterns of Implementation in a School-Wide Behavior Support Framework. *Prevention Science*, *17*(8), 992-1001.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York, NY: Macmillan.

Michie, S., Fixsen, D., Grimshaw, J. M., & Eccles, M. P. (2009). Specifying and reporting complex behaviour change interventions: the need for a scientific method. *Implementation Science*, *4*(40), 1 – 6.

Miller, F. G., & Lee, D. L. (2013). Do functional behavioral assessments improve intervention effectiveness for students diagnosed with ADHD? A single-subject meta-analysis. *Journal of Behavioral Education*, *22*(3), 253-282.

Mitchell, B. S., Adamson, R., & McKenna, J. W. (2017). Curbing our enthusiasm: An analysis of the check-in/check-out literature using the Council for Exceptional Children's evidence-based practice standards. *Behavior modification*, *41*(3), 343-367.

Mitchell, B. S., Bruhn, A. L., & Lewis, T. J. (2016). *Essential features of Tier 2 & 3 school-wide positive behavioral supports*. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), Handbook of response to intervention: The science and practice of assessment and intervention (2nd ed.). New York, NY: Springer.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical psychology review*, *11*(3), 247-266.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American journal of evaluation, 24*(3), 315-340.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology*, 267-316.

Naoom, S. F. (2014). Validation of the scores of the instructional pedagogical and instructional student engagement components of fidelity of implementation. (unpublished dissertation). University of South Florida, Tampa.

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional children*, *71*(1), 59-73.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The journal of behavioral health services & research*, *39*(4), 374-396.

Nese, R. N., Nese, J. F., McIntosh, K., Mercer, S. H., & Kittelman, A. (2017). Predicting Latency of Reaching Adequate Implementation of Tier I Schoolwide Positive Behavioral Interventions and Supports. *Journal of Positive Behavior Interventions,* 1098300718783755.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., & Duhon, G. J. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. School Psychology Review, 34, 87–106.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33-84.

Parrish, T. (2002). Racial disparities in the identification, funding, and provision of special education. In D. J. Losen & G. Orfield (Eds.), Racial inequity in special education (pp. 15–37). Cambridge, MA: Harvard Education Press.

Pas, E. T., Bradshaw, C. P., & Mitchell, M. M. (2011). Examining the validity of office discipline referrals as an indicator of student behavior problems. *Psychology in the Schools, 48*(6), 541-555.

PBISApps. (2016a, September). SWIS user's manual version 5.7. Retrieved from https://www.pbisapps.org/Resources/SWIS%20Publications/SWIS%20User%20Manual.pdf.

PBISApps. (2016b, September). CICO SWIS user's manual version 5.7. Retrieved from https://www.pbisapps.org/Resources/SWIS%20Publications/CICO-SWIS%20User%20Manual.pdf

Reinke, W. M., Herman, K. C., & Stormont, M. (2013). Classroom-level positive behavior supports in schools implementing SW-PBIS identifying areas for enhancement. *Journal of Positive Behavior Interventions*, *15*(1), 39-50.

Resnick, B., Bellg, A. J., Borrelli, B., DeFrancesco, C., Breger, R., Hecht, J., Sharp, D. L., Levesque, C., Orwig, D., Ernst, D., & Ogedegbe, G. (2005). Examples of implementation and evaluation of treatment fidelity in the BCC studies: Where we are and where we need to go. *Annals of Behavioral Medicine*, *29*(2), 46-54.

Reynolds, C. R., & Kamphaus, R. W. (2004). *BASC: Behavior Assessment Scale for Children* (2nd ed.). Circle Pines, MN: AGS Publishing.

Rindskopf, David (1984), "Structural Equation Models," Sociological Methods and Research, 13 (1), 109–119.

Roland, H. H. & Thomas, S. L., (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. (3rd ed). New York, NY: Routledge.

Ruiz-Primo, M. A. (2005). A multi-method and multi-source approach for studying fidelity of implementation. CSE: Technical Report 677. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, *38*(4), 445.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3-15.

Schaper, A., McIntosh, K., & Hoselton, R. (2016). Within-year fidelity growth of SWPBIS during installation and initial implementation. *School Psychology Quarterly, 31*(3), 358.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*(1), 32-43.

Scott, T. M., Alter, P. J., Rosenberg, M., & Borgmeier, C. (2010). Decision-making in secondary and tertiary interventions of school-wide systems of positive behavior support. *Education and Treatment of Children*, *33*(4), 513-535.

Scott, T. M., Anderson, C. M., & Spaulding, S. A. (2008). Strategies for developing and carrying out functional assessment and behavior intervention planning. *Preventing School Failure: Alternative Education for Children and Youth*, *52*(3), 39-50.

Scott, T. M., & Eber, L. (2003). Functional Assessment and Wraparound as Systemic School Processes Primary, Secondary, and Tertiary Systems Examples. *Journal of Positive Behavior Interventions*, *5*(3), 131-143.

Simonsen, B., Eber, L., Black, A. C., Sugai, G., Lewandowski, H., Sims, B., & Myers, D. (2012). Illinois statewide positive behavioral interventions and supports: Evolution and impact on student outcomes across years. *Journal of Positive Behavior Interventions*, *14*(1), 5-16.

Stefancic, A., Tsemberis, S., Messeri, P., Drake, R., & Goering, P. (2013). The Pathways Housing First fidelity scale for individuals with psychiatric disabilities. *American Journal of Psychiatric Rehabilitation*, *16*(4), 240-261.

Sugai, G., & Horner, R. H. (2009). Responsiveness-to-intervention and school-wide positive behavior supports: Integration of multi-tiered system approaches. *Exceptionality*, *17*(4), 223-237.

Sugai, G., Horner, R. H., & Lewis-Palmer, T. L. (2001). *Team Implementation checklist (TIC).* Eugene, OR: Educational and Community Supports. Available at http://www.pbis.org.

Sugai, G., Horner, R. H., & Todd, A. W. (2000). *PBIS Self-Assessment Survey 2.0.* Eugene, OR: Educational and Community Supports. Available at http://www.pbisapps.org.

Sugai, G., Lewis-Palmer, T. L., Todd, A. W., & Horner, R. H. (2001). *School-wide Evaluation Tool (SET)*. Eugene, OR: Educational and Community Supports. Available at http://www.pbisapps.org.

Sullivan, E. A. (2015). Psychometric investigation of the Check-In, Check-Out fidelity of implementation measure. (unpublished dissertation). University of Denver, Denver.

Thorndike, R. M. & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education*. (8th ed). Boston, MA: Pearson.

Turri, M. G., Mercer, S. H., McIntosh, K., Nese, R. N., Strickland-Cohen, M. K., & Hoselton, R. (2016). Examining barriers to sustained implementation of school-wide prevention practices. *Assessment for Effective Intervention*, *42*(1), 6-17.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3, 4 –70.

Vincent, C. G., Swain-Bradway, J., Tobin, T. J., & May, S. (2011). Disciplinary referrals for culturally and linguistically diverse students with and without disabilities: Patterns resulting from school-wide positive behavior support. *Exceptionality, 19*, 175-190.

Walker, B., Cheney, D., Stage, S., Blum, C., & Horner, R. H. (2005). Schoolwide screening and positive behavior supports identifying and supporting students at risk for school failure. *Journal of Positive Behavior Interventions*, *7*(4), 194-204.

Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., & Kaufman, M. J. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders, 4*, 193-256.

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. *Journal of consulting and clinical psychology*, *61*(4), 620.

Wickstrom, K. F., Jones, K. M., LaFleur, L. H., & Witt, J. C. (1996). An analysis of treatment integrity in school-based behavioral consultation. School Psychology Quarterly, 13, 141-154.

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Unpublished doctoral dissertation). University of California, Los Angeles.