AN INVESTIGATION OF THE CONTENT AND CONCURRENT VALIDITY OF THE SCHOOL-WIDE EVALUATION TOOL

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
Of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Alison E. Bloomfield
May 2015

Examining Committee Members:

Catherine Fiorello, Advisory Chair, School Psychology
Laura Pendergast, School Psychology
Joseph DuCette, Educational Psychology
Frank Farley, School & Educational Psychology
Matthew Tincani, Special Education & Applied Behavior Analysis

ABSTRACT

The School-wide Evaluation Tool (SET) is a commonly used measure of the implementation fidelity of school-wide positive behavior interventions and supports (SWPBIS) programs. The current study examines the content and concurrent validity of the SET to establish whether an alternative approach to weighting and scoring the SET might provide a more accurate assessment of SWPBIS implementation fidelity. Twenty published experts in the field of SWPBIS completed online surveys to obtain ratings of the relative importance of each item on the SET to sustainable SWPBIS implementation. Using the experts' mean ratings, four novel SET scoring approaches were developed: unweighted, reweighted using mean ratings, unweighted dropping lowest quartile items, and reweighted dropping lowest quartile items. SET 2.1 data from 1,018 schools were used to compare the four novel and two established SET scoring methods and examine their concurrent validity with the Team Implementation Checklist 3.1 (TIC; across a subsample of 492 schools). Correlational data indicated that the two novel SET scoring methods with dropped items were both significantly stronger predictors of TIC scores than the established SET scoring methods. Continuous SET scoring methods have greater concurrent validity with the TIC overall score and greater sensitivity than the dichotomous SET 80/80 Criterion. Based on the equivalent concurrent validity of the unweighted SET with dropped items and the reweighted SET with dropped items compared to the TIC, this study recommends that the unweighted SET with dropped items be used by schools and researchers to obtain a more cohesive and prioritized set of SWPBIS elements than the existing or other SET scoring methods developed in this study.

I dedicate this dissertation to the memory of my father, Donald Bloomfield, whose unconditional love and support continues to inspire me every day. His commitment to family and intellectual curiosity above all else have helped me to reach this milestone and have served as a guidepost for all personal and professional achievements in my life.

ACKNOWLEDGMENTS

I am sincerely grateful for all of the help, support, and encouragement I have received from my advisor, committee, colleagues, friends, and family throughout not only this dissertation process, but also the path through my doctoral program. Most of all, I cannot give enough thanks to Dr. Catherine Fiorello, who has kept me going and encouraged me throughout over the past six years. She has provided me with constant guidance and perspective through the ups and downs of the program and dissertation, and I cannot overstate how much her confidence in me has meant.

I would also like to extend my thanks to the other members of my dissertation committee, Dr. Laura Pendergast, Dr. Joseph DuCette, Dr. Frank Farley, and Dr. Matthew Tincani, for their insight, time, comments, and suggestions.

This project would not have been possible without the help and support of Dr. Robert Horner, Dr. Kent McIntosh, and Robert Hoselton at the Educational and Community Supports research program at the University of Oregon. I am grateful for their help in refining my research ideas and their support for my project. Most especially, I thank them for believing in my project enough to allow me to use their database.

Finally, I would like to thank all those who have personally supported me over the past six years. To my mother and brother, thank you for loving me unconditionally, encouraging me, and being my biggest cheerleaders. Last, but certainly not least, I would like to thank my husband, John Graettinger, for his love, support, late night pep talks, and for helping me to celebrate even the smallest of victories. I cannot thank everyone enough for giving me all I needed to achieve this milestone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The United States education system faces the challenge of serving the complex academic and behavioral needs of students from a wide range of cultural, racial, and socioeconomic backgrounds. Despite increased emphasis on responsibly serving the needs of students from culturally and linguistically diverse populations (e.g., considerations for student language and cultural background outlined in the Individuals with Disabilities Education Act [IDEA] of 2004) and accountability for student outcomes, there remain discrepancies in the experiences of students related to their cultural or linguistic backgrounds. When comparing students from different cultural, racial, or ethnic backgrounds, there is variability in how student behavior problems are handled (Skiba, Michael, Nardo, & Peterson, 2002) and in rates of classification of students as having emotional disturbances (U.S. Department of Education, 2011; National Education Association [NEA], 2007). These issues of disproportionality in responses to student behavior and behavior-based special education classification signal the need for educational reform, including a prevention- and early intervention-based approach to student behavior and greater accountability for fairness in disciplinary practices. With the 2004 reauthorization of IDEA, school districts were permitted to begin using up to 15% of their federal special education funding to develop prevention and early intervention programs such as academics-targeted response to intervention (RtI; see Definition of Terms at end of Chapter 1) or school-wide positive behavior interventions and supports (SWPBIS; see Definition of Terms). Therefore, federal

education legislation has moved toward the development of problem-solving prevention models for both academics and behavior, with increased expectations for data-based decision making and accountability for demonstrating the effectiveness of these programs.

SWPBIS has been promoted and has shown promise as a system-level problem-solving model aimed at decreasing disproportionality and providing universal procedures and supports for teaching and responding to students' behaviors. As early as the 1980s, researchers in school psychology had begun to describe efforts to implement school-wide prevention activities with the goal of decreasing numbers of referrals for special education, specifically through the use of pre-referral intervention and consultation (e.g., Ponti, Zins, & Graden, 1988). Though general frameworks based in consultation were proposed (Ponti et al., 1988), with Effective Behavior Support (EBS [see Definition of Terms at end of Chapter 1]; Walker, Horner, Sugai, & Bullis, 1996; Lewis & Sugai, 1999) and SWPBIS (Sugai, Horner et al., 2000) the field moved toward specific sets of procedures for the establishment of school- and district-wide reforms in how schools approach behavior. Furthermore, positive behavior interventions and supports (PBIS; see Definition of Terms) is the only behavior intervention explicitly mentioned in IDEA or any other federal education law (Office of Special Education Programs [OSEP] Technical Assistance Center on PBIS, 2013b). As of 2007, data collected through the OSEP Technical Assistance Center on PBIS and the National Center for Education Statistics (NCES) indicated that SWPBIS was being implemented in 7,953 schools across 47 states (including Washington, D.C.), even if some states had only just begun implementation with a few schools (Spaulding, Horner, May, & Vincent, 2008). In October 2013, the

OSEP Technical Assistance Center on PBIS estimated that the number of schools implementing SWPBIS had grown to approximately 18,275 (2013a). Through its alignment with education legislation and research documenting its effectiveness and promise (e.g., Chitiyo, May, & Chitiyo, 2012; Horner, Sugai, & Anderson, 2010; Horner, Sugai, Smolkowski, Eber, Nakasato et al., 2009), SWPBIS has become increasingly popular and ubiquitous in schools across the United States.

Research over the past two decades has demonstrated that implementation of SWPBIS is related to improvements in disproportionality of disciplinary responses. For example, Tobin and Vincent (2011) found that improved SWPBIS implementation fidelity was associated with decreases in unfair or disproportionate exclusionary punishments. However, another study published by Vincent and Tobin (2011) found that decreases in exclusionary punishments associated with SWPBIS were greatest for White students, and African American students continued to experience disproportionate levels of exclusionary discipline practices. In addition, SWPBIS implementation has been shown to effect improvements in student behaviors (Solomon, Klein, Hintze, Cressey, & Peller, 2012), teacher-reported well-being (Ross, Romer, & Horner, 2012), and even student academic performance (Horner et al., 2009). Though individual research groups often conclude that SWPBIS is associated with positive effects on disproportionality and student/school outcomes, recent meta-analyses of SWPBIS research have indicated, overall, that the use of implementation fidelity assessment in the SWPBIS literature base tends to be lacking both in how frequently it is occurring (Solomon et al., 2012) and in researchers' adherence to established standards for successful implementation (Chitiyo et al., 2012).

In SWPBIS research, it is critical that researchers use measures of SWPBIS implementation fidelity to establish that the successes noted in these studies are related to SWPBIS and not other school factors or changes in policy. These measures ensure that the SWPBIS framework and procedures were followed and developed appropriately. Both IDEA 2004 and leaders of the field of school psychology have emphasized the importance and necessity of selecting evidence-based practices and interventions as well as research-supported curricula for use in both general education and special education settings. Kratochwill and Shernoff (2004), along with the Task Force for Evidence-Based Interventions in School Psychology, have developed strategies for implementing evidence-based practices in school, including suggesting the use of manuals and procedural guidelines that allow for flexibility to fit the specific context and needs of a school (p. 37) and performing checks on adherence to implementation procedures. In practice, implementation fidelity measurement also has the additional benefit of allowing schools to progress monitor the development of their SWPBIS plan and develop an action plan for continued refinement and building local capacity for sustainability.

SWPBIS implementation may be measured at several levels (e.g., universal prevention supports or more targeted intervention supports) and through different methods such as outside evaluation or school-based team self-report. In evaluating the implementation fidelity of the core components of SWPBIS intended to provide universal behavior education and supports to all students there are three commonly-used measures: the School-wide Evaluation Tool (SET; Version 2.1; Sugai, Lewis-Palmer, Todd, & Horner, 2005), the Team Implementation Checklist (TIC; Version 3.1; Sugai, Horner, Lewis-Palmer, & Rossetto Dickey, 2011), and the School-wide Benchmarks of Quality

(BoQ; Revised; Kincaid, Childs, & George, 2010). Of particular relevance to this study is the SET, which was developed in 2001 (Sugai, Lewis-Palmer, Todd, & Horner) as a tool for researchers to evaluate the implementation level of SWPBIS in schools (Horner, Todd, Lewis-Palmer, Irvin, Sugai, et al., 2004). The SET is a 28-item measure that assesses implementation across seven categories of SWPBIS plan components, including: Expectations Defined; Behavioral Expectations Taught; On-going System for Rewarding Behavioral Expectations; System for Responding to Behavioral Violations; Monitoring and Decision-Making; Management; and District-Level Support. See Appendix A for a copy of the SET Version 2.1; hereafter "SET" refers to Version 2.1 unless otherwise noted. The SET was designed as a measure of SWPBIS implementation fidelity that could be efficiently completed in approximately two hours by evaluators who were not affiliated with the school being evaluated (Horner et al., 2004). Studies focused on the validation of the SET (Horner et al., 2004; Vincent, Spaulding, & Tobin, 2010) and comparison of the SET against other measures of SWPBIS implementation (Cohen, Kincaid, & Childs, 2007) have raised questions regarding the validity of the scores obtained through the SET including the degree to which its current scoring methods reflect those aspects of SWPBIS implementation that are most critical for effective and sustainable implementation.

Though SWPBIS evolved out of concerns regarding disproportionality, issues of cultural fairness and cultural sensitivity within SWPBIS research have gotten lost as the field works to establish the efficacy and evidence-base for SPWBIS. So while SWPBIS has been shown to be a promising means of decreasing that disproportionality and improving student educational experiences, issues of implementation fidelity

measurement and the consistency with which those measures are used must first be resolved. Therefore, since it is established that implementation fidelity is a critical part of research and critical to building SWPBIS and achieving related outcomes, there remains a need for continued examination and validation of tools for measuring SWPBIS, specifically the SET. The improvement and continued refinement of SWPBIS implementation fidelity measures will allow future research to more conclusively and confidently investigate the outcomes of SWPBIS implementation. The current study seeks to evaluate and re-examine the content and concurrent validity of the SET by addressing the following research questions:

1. Considering the existing items assessed by the SET, which items/factors do published experts in the field of SWPBIS feel are the most critical for successful implementation of SWPBIS? Do the experts' ratings suggest an alternative method of scoring the SET that might improve its ability to measure those factors that are most important?
2. When the SET is rescored based on expert input (using item weightings and/or removal of unimportant items), does that improve the SET's accuracy in measuring implementation and improve its concurrent validity with a secondary measure of SWPBIS implementation?

Definition of Terms

Applied Behavior Analysis (ABA) – Applied behavior analysis is the application of behavioral principals to the development of socially important behaviors (Wolf, 1978) that improve an individual's quality of life. This is achieved through an analysis of the function of the individual's behavior and developing environmental structure and contingencies to increase desired behaviors and minimize disruptive, self-injurious, or dangerous undesired behaviors (refer to Baer, Wolf, and Risley, 1968).

Positive Behavior (Interventions and) Support (PBS; PBIS) – Referred to as both PBS and PBIS, positive behavior interventions and supports is the application of applied behavior analytic techniques in a person-centered approach to modify an individual's environment to "enhance quality of life and minimize problem behaviors" (Carr, Dunlap, Horner, Koegel, Turnbull, et al., 2002; p. 4). This approach emphasizes the use of reinforcement to encourage positive behaviors (rather than punishment of undesired or maladaptive behaviors), manipulating behavioral antecedents, and teaching effective replacement behaviors. PBIS is the only behavioral intervention explicitly mentioned in federal education law (OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports, 2013b).

Effective Behavior Support (EBS) – Effective behavior support was one of the first names developed for school-wide positive behavior support programs (Lewis & Sugai, 1999), becoming more uniformly referred to in the literature as school-wide positive behavior interventions and supports. The *Effective Behavior Support Survey (EBS Survey*; Sugai, Todd, & Horner, 2000) is still used to assess staff-reported SWPBIS implementation.

Response to Intervention (RtI) – Rather than referring to a specific curriculum or program, response-to-intervention (RtI) is a framework for establishing a multi-tiered system of student instruction, assessment, and supports. Key features of RtI include systematic programming as well as data-based decision making and problem solving (Brown-Chidsey & Steege, 2005). Within an RtI system, all students are provided instruction using evidence-based curricula, progress is monitored to measure student learning, and students struggling within the curricula are provided group-based or individual support and intervention. RtI tiered support frameworks may be applied to both academics and student behavior (see SWPBIS).

School-wide Positive Behavior Interventions and Supports (SWPBIS) – School-wide positive behavior interventions and supports (referred to both with and without "interventions" in the title, and abbreviated as SWPBIS, SWPBS, and SW-PBIS) is the systemic application of PBIS at a school-wide level. SWPBIS systems integrate PBIS and a tiered RtI approach to teaching and maintaining behavioral rules and expectations. Key features of SWPBIS include universal behavioral education programs, systems of reinforcement, consistent consequences for behavioral violations, data collection, data-based decision making, and tiered supports/interventions (Lewis & Sugai, 1999; Sugai, Horner, et al., 2000). See Chapter 2 for a comprehensive discussion of SWPBIS.

CHAPTER 2

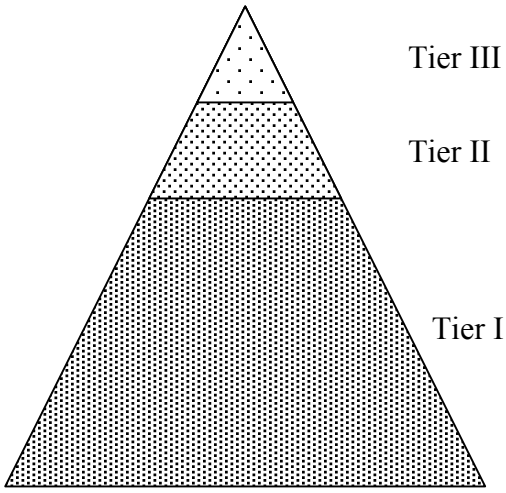
REVIEW OF LITERATURE

School-wide Positive Behavior Interventions and Support

Education and the field of school psychology are moving away from reactionary "wait-to fail" models based on referrals toward proactive early identification and intervention models (Reschly, 2004; Shapiro, 2000; IDEA, 2004). Response to intervention (RtI) frameworks for both academics and behavior have emerged as objective problem-solving and data-based service delivery systems (Brown-Chidsey & Steege, 2005). For teaching and managing student behaviors using response-to-intervention, SWPBIS has developed as a set of procedures that applies behavioral principles of reinforcement and environment-based contingencies to school behavior policies. SWPBIS focuses on creating universal behavior guidelines and tiers of positive behavior supports and interventions (Lewis & Sugai, 1999; Sugai, Horner, et al., 2000). Aimed at supporting all students, the universal supports in a SWPBIS system include a set of three to five positively-stated school rules, clear behavioral expectations for each specific area of the school, a behavior curriculum (plan for teaching students the rules/expectations), a system of reinforcing appropriate student and staff behavior, as well as transgression-specific corrective consequences (Lewis & Sugai, 1999). Carr et al. (2002) described how the use of PBIS had evolved out of effective use in populations of individuals with cognitive and developmental disabilities, stemming from foundations in applied behavior analysis (ABA; see Definition of Terms). SWPBIS elevates PBIS from an individualized to a systems-level approach, maintaining its goal to "enhance quality of

life and minimize problem behaviors" (Carr et al., 2002, p. 4). Sugai, Horner, and colleagues (2000) describe SWPBIS as a system based on measureable behaviors, data-based decision making, the use of evidence-based interventions and practices, as well as systemic change to increase sustainability. The theoretical bases for SWPBIS include prevention science, applied behavior analytic techniques, an instructional approach to teaching appropriate behaviors, integration of evidence-based practices and interventions at every level of support, and a systems-level approach to sustainable, long-term change (Sugai, Horner, & McIntosh, 2007).

SWPBIS is conceptualized as a behavior-focused counterpart to academics-focused RtI frameworks, both of which are commonly presented visually as a three-tiered triangle. A typical visual representation of SWPBIS is presented in Figure 1, in which the students move to higher tiers as it is determined that they need more intensive or individualized levels of support.



Tier III – Tertiary Interventions
(Individualized Interventions;
Approximately 5% of students)

Tier II – Secondary Prevention/Intervention
(Small Group Interventions;
Approximately 15% of students)

Tier I – Primary Prevention / Universal Supports
(Approximately 85% of students)

Note: Triangle sections not drawn to scale; Based on service-delivery model described in SWPBIS literature. For example, see Sugai, Horner, et al. (2000) and OSEP Technical Assistance Center on Positive Behavioral Interventions & Supports (2009b; 2013a).

Figure 1. Three-Tiered SWPBIS Framework

At Tier I, SWPBIS provides primary prevention in the form of a universal behavior support system, with specific school-wide and classroom-based behavior teaching as well as consistent procedures for reinforcement and behavioral corrections. Though the concept of teaching students appropriate behaviors and coping strategies is not novel, the SWPBIS framework differs from commercially-available behavior education or social skills programs, in its emphasis on utilizing applied behavior analytic approaches, fitting the context and needs of a specific school, and its multi-tiered approach to responding to students who struggle to meet behavior expectations. This approach is unique in that SWPBIS is not a specific, canned curriculum or program, but rather a set of procedures and a framework that enables a school, through the support of an expert consultant, to develop appropriate applied behavior analytic practices for the needs of their students, school, and community (OSEP Center on Positive Behavioral Interventions and Supports, 2009b). According to the model presented by Sugai, Horner, and colleagues (2000), when SWPBIS is implemented with fidelity, approximately 85% of students respond to primary prevention supports at Tier I, but the 15% of students who continue to struggle move on to the Tier II – secondary prevention with small group intervention. At Tier III, students who are still demonstrating persistent behavior problems after receiving small group interventions (approximately 5%) are evaluated using a functional behavioral assessment and individualized behavioral interventions and modification programs. This tiered system allows schools to allocate support resources to those students who need them based on student response to behavioral instruction and intervention.

The Office of Special Education Programs (OSEP) within the United States Department of Education provides training and resources for schools and districts

beginning the process of SWPBIS development through its Technical Assistance Center on PBIS (www.pbis.org, 2013a). Whereas many academic and behavioral intervention programs may be purchased and implemented with only minimal training, the SWPBIS implementation blueprint describes how the development of a SWPBIS system must be school-specific and involves both the commitment and involvement of administrators and the involvement of consultants to guide the process (Sugai et al., 2010). Given that SWPBIS is a framework rather than a specific intervention, consultants work with administrators and behavior support teams or PBIS teams composed of a representative sample of faculty to help the school develop rules, plans, reinforcement systems, and consequence systems to fit the specific needs and culture of a school.

SWPBIS Tiered Service Delivery

Schools embarking on the development of a SWPBIS program begin with the development of Tier I supports and the underlying structure that will allow for data-based decision making and appropriate service delivery across all levels of implementation (Sugai et al., 2010). Universal prevention and supports at Tier I include the development of three to five positively stated school rules, defining those rules as location-specific behavior expectations, developing a standardized behavior education program that targets improvement and maintenance of the behavior of the students and school as a whole, as well as consistent systems for rewarding behavior, corrective consequences, and data collection and use (Horner, Sugai, Todd, & Lewis-Palmer, 2005; Lewis & Sugai, 1999). Turnbull and colleagues (2002) describe the universal component of SWPBIS as being, "proactive in that every student gets effective PBS without identification or referral" (p. 380) and targeted at addressing the needs of the school as a whole. The development of

these initial components of SWPBIS requires the commitment of leadership within the school, development of systems for administrative accountability, staff training, staff buy-in, professional development for staff and administrators, as well as monitoring of implementation through fidelity assessment (Handler et al., 2007).

At the core of Tier I development is the premise that schools cannot and should not assume that their students arrive at the school doors having already learned the difference between "good" and "bad" behaviors, but rather that behavior education must be incorporated into the school's core instructional practices (Sugai et al., 2007). Cartledge and Milburn (1995; as cited in Cartledge, Tillman & Johnson, 2001) describe a general model for teaching students appropriate school behaviors. Within their framework, similar to a hierarchy for learning any academic skill (Daly, Lentz, & Boyer, 1996), students are instructed in expected behaviors, those behaviors are modeled for them, the students are provided with frequent opportunities to practice and receive reinforcement or feedback on those behaviors, and the students explicitly aided in maintaining and generalizing those skills. At Tier I (and subsequent tiers), SWPBIS requires a high level of systems-level, administrative, team-based, and classroom-level consultation (see Erchul & Martens, 2002) to guide the development of supports, provide training and assessment, and guide data analysis and problem solving processes.

A critical component of SWPBIS Tier I implementation is the development of a systematic method of collecting data on student behavior transgressions. While it is possible to use direct observation of student behavior to obtain frequency counts of inappropriate behaviors within a classroom or specific school location, it is impractical to use systematic direct observation to capture the behaviors of the school as a whole for

adequate decision making. Therefore a core component of SWPBIS development and implementation is the use of office discipline referrals (ODR). ODRs are forms commonly used by schools to record student behavioral infractions and to provide documentation of rule violations and how they are handled by teachers or administrators. During SWPBIS development, schools are guided through procedures related to the development of standardized ODR forms and how they are used. This standardization is critical as ODRs are not only a key component of school-wide evaluation practices and data-based decision making within SWPBIS (Turnbull et al., 2002), but also serve as an outcome measure to evaluate the effectiveness of the SWPBIS plan. The collection and summarizing of ODRs provides an indirect measure of student behavior, allowing schools to track behavioral infractions by many factors including: student, frequency, type of infraction, time, location, consequences, and possible motivations for the behavior. Though most schools track behavioral infractions, there is substantial variability in what types of information are collected, what behaviors warrant an ODR, and how they are tracked (Sugai, Sprague, Horner, & Walker, 2000). Therefore, SWPBIS implementation procedures include the development of standardized data collection procedures for how and when to record rule-violating behaviors as well as for the adoption of a system of entering and analyzing data, such as the School-Wide Information System (SWIS; May et al., 2006). Research over the past ten years has found ODRs to be a meaningful indicator of students requiring Tier II or Tier III support within schools implementing SWPBIS (Irwin, Tobin, Sprague, Sugai, & Vincent, 2004; McIntosh, Campbell, Carter, & Zumbo, 2009). For example, McIntosh and colleagues (2009) examined the concurrent validity of ODRs with the Behavior Assessment System

for Children (BASC-2, Reynolds & Kamphaus, 2006), finding a significant relationship between elevated numbers of ODRs and clinically elevated scores on the BASC-2 Externalizing Composite. In other words, it is possible to use ODR cut scores as a screening tool for identifying students with clinically elevated frequencies of externalizing behaviors such as hyperactivity or rule-breaking behaviors, though it is more difficult to identifying students at-risk for depression or anxiety with ODRs. These results support the use of ODRs as a tool for measuring overall student responsiveness to SWPBIS, but only once schools have implemented SWPBIS with fidelity and adherence to procedures for the completion and use of ODR forms and only providing information regarding externalizing behaviors.

At Tier II, just as within an RtI framework targeting academic behavior and achievement (Brown-Chidsey & Steege, 2005), supports are developed to serve those students who have not responded to the school's universal supports. Approximately 15% of students continue to receive elevated frequencies of ODRs and fail to respond adequately to the universal supports at Tier I (Lewis & Sugai, 1999; Sugai et al., 2010). The goal of Tier II interventions is to reduce "the number of existing cases of problem behaviors by establishing efficient and rapid responses to problem behavior" (Sugai et al., 2010). Once a school has established implementation at Tier I, group programs are established at Tier II to provide additional support and opportunities for reinforcement. Depending on the needs of students, these may include social skills groups, group counseling, behavior education programs such as Check in/Check out (Filter et al., 2007), or standardized daily report cards.

The third tier of SWPBIS systems involves the most intensive and individualized supports for those students who continue to struggle to meet the behavioral expectations established at the universal level, even after intensive group-based intervention at Tier II (Lewis & Sugai, 1999; Sugai, Horner, et al., 2000). The *SWPBS Implementation Blueprint* (Sugai et al., 2010) describes Tier III as focused on "reducing the intensity and/or complexity of existing cases of problem behavior that are resistant to primary and secondary prevention efforts" (p. 20). At Tier III, students who continue to struggle are assessed using a functional behavior assessment to examine the function of their behaviors and provide individualized contingencies to decrease the frequency and intensity of those behaviors (O'Neill et al., 1997). It is estimated that approximately 5-7% or less of the student population in a typical school would require use of the tertiary intervention services (Lewis & Sugai, 1999; Sugai et al., 2000).

SWPBIS Efficacy

SWPBIS has been demonstrated to be effective at modifying and improving student behaviors through a number of single-case design studies as well as through small-scale randomized controlled trials. While much research on SWPBIS focuses on the entire school as the unit of study, examining behavior rates or trends for the students of that school in aggregate, studies have also shown the effectiveness of SWPBIS's data-based approach to problem-solving and team-based decisions by empirically approaching location-specific problem behaviors. Studies have found that schools using systematic data collection and team-based analysis of those data through a system such as SWIS (May et al., 2006) have had success at identifying and intervening in location-specific behavior issues as well as specific behavioral problems within the school. Success has

been found in decreasing rates of problem behavior on school busses (Putnam, Handler, Ramirez-Platt, & Luiselli, 2003), decreasing problem behaviors on the playground (Lewis, Powers, Kelk, & Newcomer, 2002), and even decreasing the decibel level in school hallways during transitions (Kartub, Taylor-Greene, March, & Horner, 2000). SWPBIS has also been found to have additional benefits to school functioning with regard to time as a valuable resource for teachers and administrators. Scott and Barrett (2004) found that within schools implementing a SWPBIS program with fidelity, decreases in the number of students referred to the office are associated with administrators spending less time managing behaviors and teachers regaining instructional time in their classrooms that might otherwise have been dedicated to filling out ODR forms.

The efficacy of SWPBIS has also been studied through randomized and wait-list controlled trials. Horner and colleagues (2009) studied the implementation progress and associated outcomes across 30 elementary schools in Hawaii and 30 elementary schools in Illinois, half of which were randomly assigned to receive state-provided SWPBIS training and support immediately and half of which were assigned to begin training a year later. Each school was assessed three times: before receiving any training, between when the treatment group began training and when the control group began training, and after both groups had received at least a year of training. At each point, the researchers assessed SWPBIS implementation fidelity using the SET (Sugai, Lewis-Palmer, et al., 2001), perceived school safety, office discipline referral rates, and standardized state academic test scores. Looking at SWPBIS implementation via total scores obtained on the SET, the researchers found that both the immediate treatment group and waitlisted

group showed significant improvement on the SET from before receiving SWPBIS training to after training. Prior to training, the schools had access to only SWPBIS training materials, but no direct consultation or training from their state PBIS initiatives. Additional analyses found that SWPBIS training was associated with an improvement in perceived school safety, as well as preliminary indications that continued SWPBIS implementation might support improved academic performance on state accountability assessments.

Bradshaw, Mitchell, and Leaf (2010) also conducted a randomized-controlled trial of the effects of SWPBIS on student and school outcomes at the elementary level. This study focused on the implementation of SWPBIS in 37 elementary schools across Maryland, drawing from five school districts form rural or suburban areas. Of the 37 schools that volunteered to participate in the study, 21 were randomly assigned to the "SWPBIS" condition and underwent SWPBIS training and coaching, while 16 were randomly assigned to the "Comparison" condition and their administrators signed an agreement that they would not seek to implement SWPBIS during the research period. Schools in the SWPBIS condition established teams, attended 2-day training sessions and annual booster sessions, were provided a coach, and were guided through the process of developing SWPBIS plans, implementation action plans, and professional development programs. SWPBIS implementation fidelity was evaluated using both the SET and the EBS Survey (Sugai, Todd, et al., 2000) and outcome data were gathered through ODRs, school suspension rates, and state standardized assessment results. Across five years of implementation, the researchers found large effect sizes for a statistically significant interaction effect of the SWPBIS implementation training process over time on overall

SET scores ($p < .001$, $d = 3.22$) as well as on EBS Survey subscales, including having schoolwide systems in place ($p < .001$, $d = 1.71$). These findings demonstrate that the SWPBIS training and coaching procedures utilized within this study (based on typical training procedures utilized by Maryland's state initiative) are effective at establishing and maintaining the core components of SWPBIS. Moreover, the schools implementing SWPBIS demonstrated statistically significant decreases in the percentage of students receiving an ODR and the number of ODRs per student, and showed a small effect size for a statistically significant decrease in suspension rates compared to schools in the comparison condition ($p = .03$, $d = .27$). Though nonsignificant, trends for improvement in academic achievement were greater in schools implementing SWPBIS than those not implementing SWPBIS. This study contributes to the evidence for the efficacy of SWPBIS by demonstrating that schools implementing SWPBIS with fidelity through coaching and training demonstrate reductions in ODRs as well as decreases in suspension rates compared to schools without SWPBIS.

Evaluating the Efficacy of SWPBIS in the Aggregate

After almost two decades of SWPBIS implementation, development, and studies, researchers have begun to look at the outcomes of SWPBIS systems in the aggregate to examine the overall impact of SWPBIS. Solomon et al. (2012) conducted a meta-analysis of single-subject/case design studies that had looked quantitatively at the effectiveness of SWPBIS on single schools, school-specific locations (e.g., classroom or hallway), or groups of schools in a district. In the case of each of the twenty studies evaluated by Solomon and colleagues, the school or schools were considered to be a single unit of study. Outcome measures used by these studies included direct

observations (frequencies) of problem behavior, ODRs, academic achievement, and types/frequencies of disciplinary consequences. The meta-analysis found that SWPBIS had been moderately effective in decreasing rates of problem behavior across multiple tiers, with increasing effectiveness at Tier III in which students received individualized behavior plans or accommodations. Effects on ODRs were less pronounced, which the authors attribute to referrals being an indirect, less reliable, and more variably used measure. The researchers noted, however, that while treatment fidelity is acknowledged to be a critical element of any evidence-based intervention program, only twelve of the 20 studies reviewed measured the implementation fidelity of the program being implemented, including through the use of an implementation checklist or survey/assessment by a consultant through the SET.

In another meta-analysis and review completed by Chitiyo et al. (2012), the researchers focused on the identification and study of experimental SWPBIS research designs, rather than single-subject designs. In this study, the authors identified ten experimental designs; however, they noted that only two of the studies met the standards of methodological rigor necessary to be considered empirical evidence of SWPBIS' efficacy. Seven out of the ten studies reviewed demonstrated positive effects, including decreased ODRs, improvements in grades or on standardized state examinations, fewer suspensions, and improved observed student behavior. While Solomon and colleagues (2012) had found that most single-case studies they evaluated had not used any appropriate measure of SWPBIS implementation fidelity, this study found that eight of the ten experimental designs used acceptable measures of implementation fidelity. However, though fidelity was measured and reported, only two of the ten schools

reported implementation levels consistent with high fidelity standards (usually 70-80% depending on the measure utilized). Based on this finding, researchers are utilizing more implementation fidelity measures; however, they are not holding their schools to high fidelity standards when considering outcomes associated with SWPBIS.

Based upon both recent meta-analyses of SWPBIS research, it is evident that despite the existence of measures of SWPBIS implementation fidelity and noted improvement in their frequency of use, they are not being consistently used to determine level of adherence to SWPBIS procedures and components. Interestingly, though implementation fidelity is emphasized as a critical component of SWPBIS and any evidence-based intervention (IDEA 2004; Cooper, Heron, & Heward, 2007; Kratochwill & Shernoff, 2004), it is not always examined within studies of SWPBIS, making it difficult to know how much of the observed changes might be due to SWPBIS procedures rather than other school variables or programs.

SWPBIS Efficacy in Decreasing Disproportionality

As evidenced by the previous paragraphs, much of the research on SWPBIS efficacy and outcomes has been focused on general student behavior trends, academic outcomes, and school resources. However, despite the fact that SWPBIS evolved out of a need for greater fairness in behavior education and disciplinary procedures (e.g., Lewis & Sugai, 1999; Ponti et al., 1988; Walker et al., 1996) there has been limited investigation into the effectiveness of SWPBIS as a means of decreasing disproportionality. Considering disproportionality in disciplinary actions taken against children from African American backgrounds, implementation of SWPBIS with fidelity has been associated with a significant decrease in exclusionary punishments such as expulsions and

suspensions (Horner et al., 2005). Researchers have continued this line of research, looking more specifically at the relationship between SWPBIS implementation and the use of exclusionary disciplinary practices with students from diverse ethnic backgrounds.

In a study of 46 schools, Tobin and Vincent (2011) examined the relationship between SWPBIS implementation and disproportional exclusions of African American students. The authors measured exclusions as ODRs resulting in suspension, expulsion, or otherwise removing the student from instructional and social settings within the school. The researchers compared suspension and expulsion data gathered through the SWIS database (May et al., 2006) to SWPBIS implementation level as measured by the EBS Survey (Sugai, Todd, et al., 2000). This SWPBIS implementation fidelity tool evaluates Tier I disciplinary procedures, implementation at the school-wide level across both classroom and nonclassroom settings, classroom behavior management systems, as well as more targeted systems for addressing students at risk for severe behavior problems. Tobin and Vincent (2011) utilized a relative rate index (RRI) as a measure of disproportionality, which compares the number of exclusions per 100 students by dividing the number of exclusions by the total number of students enrolled, and dividing the rate for African American students by the rate for white students. Overall, the RRI for the sample indicated that African American students were 3.11 times more likely to receive an exclusionary punishment than White students, despite the fact that all of the included schools had been implementing SWPBIS throughout the study. Improvements on specific items on the EBS Survey, "Expected student behaviors are acknowledged regularly (positively reinforced) (> 4 positives to 1 negative)" ($\beta = -.812$; $p = .003$) and "Transitions between instructional and non-instructional activities are efficient and

orderly" (β = -.606; p = .014) were more highly related to statistically significant decreases in RRI. It is worth noting, however, that using increases on individual items to demonstrate improved implementation is not a robust method of measuring overall SWPBIS implementation fidelity.

Within the Tobin and Vincent (2011) study, among the thirteen schools that improved their EBS Survey scores between across two school years, eight demonstrated a decrease in the RRI associated with improved implementation fidelity – thereby indicating a decrease in disproportional exclusionary punishments given to African American students. Regression analyses indicated that decreases in these eight schools were associated with the use of data-based decision making, ongoing training, and the use of functional behavior assessments and behavior plans. This study therefore indicates that improvements in specific areas of SWPBIS implementation fidelity are associated with decreases in disproportionate exclusions, particularly in relation to specific implementation factors. Moreover, Tobin and Vincent (2011) point out that the use of implementation fidelity assessment can help the field identify specific strategies that may decrease disproportionality. Replication of this type of research with a greater number of schools is necessary to demonstrate decreases in disproportional exclusions with greater power. Furthermore, the use of an implementation fidelity tool such as the SET would allow for a potentially more objective measure of SWPBIS implementation, compared to the EBS Survey, which is completed by cumulative responses from school staff.

Vincent and Tobin (2011) more broadly investigated disproportionality, looking at whether decreases in out-of-school suspensions equally affected students across ethnic backgrounds and disability status. The study found that students from African American

backgrounds appeared to benefit less than their white peers from decreases in exclusionary discipline practices associated with increases in SWPBIS implementation; however, the results were based on questionable foundations. Similar to their study discussed above (Tobin & Vincent, 2011), Vincent and Tobin used the EBS Survey (Sugai, Todd, et al., 2000), ODR, and student disability status obtained through pbssurveys.org (now PBISapps.org) and the SWIS database (May et al., 2006), supplemented with student ethnicity data from the NCES. The EBS Survey, however lacks a "validated criterion score associated with full SWPBS implementation" (Vincent & Tobin, 2011, p. 219), and therefore the researchers based implementation changes on difference scores calculated subtracting the EBS Survey subscale scores (reflecting implementation at the schoolwide, classroom, nonclassroom, and individual support levels) across two years. Using these difference scores, the researchers concluded that overall the schools had demonstrated improved SWPBIS implementation, but failed to provide analyses that proved statistical significance. Therefore, while they were able to analyze out-of-school suspension rates across ethnicity and disability groups, the study failed to establish a valid basis for attributing changes to SWPBIS implementation levels.

Overall, though SWPBIS is often recommended as a means of decreasing disproportionality in special education identification and exclusionary discipline practices (e.g., NEA [2007]; Sprague, Vincent, Tobin & CHiXapkaid [2012]), the field is in need of more comprehensive studies. Specifically, these studies must compare changes in disproportionality with validated and reliable measures of SWPBIS implementation fidelity. Without these validated measures and scores that truly reflect SWPBIS

implementation, researchers will not be able to confidently link SWPBIS with the desired behavioral and academic outcomes.

SWPBIS Implementation

Implementation Theory and Systems-Level Change

As discussed above, as educational and school-based mental health regulations and policies have required schools to adopt evidence-based practices, implementation fidelity and effectively measuring implementation have become increasingly important. Particularly in the case of RtI and SWPBIS, which require that states and districts adjust federal funding streams for special education and intervention programs (IDEA 2004), it is crucial that there be valid and reliable tools for measuring the degree to which districts are correctly and effectively translating and applying research to practice. Fixsen, Blase, Naoom, and Wallace (2009) describe how this movement in human service fields like education has become an "active process" in which evidence-based interventions are not a canned or pre-determined product, but rather that the practitioner's use of the intervention is what determines its effectiveness. Therefore, Fixsen and colleagues (2009) propose that the use of evidence-based practice implementation on larger scales may necessitate involving the help of expert consultants to support and guide practitioners "to achieve high fidelity use of the products of science and to assure benefits to its customers" (p. 532).

Based on a review of research to practice literature, Fixsen et al. (2009) developed a theoretical framework of implementation, "Core Implementation Components," in which specific factors drive and support high-fidelity implementation behaviors of practitioners. These core implementation components include: staff selection, training

before and throughout implementation, ongoing coaching and consultation, evaluation of staff performance/behaviors, a system to support decision-making, and administrative support (Fixsen et al., 2009). The authors suggest that it is more helpful to look at the combination of these components or factors influencing implementation fidelity than it is to consider implementation phases or individual factors. This approach to attaining sustainable implementation through multiple intersecting and interacting components is shared and mirrored through the work of the National Technical Assistance Center on PBIS and prominent theorists and researchers in the field. They present a model for building capacity and sustainable implementation through the constant practices of assessing and prioritizing outcomes, continuously engaging in forms of self-assessment through data collection and analysis, engaging in evidence-based practices, and practicing implementation with fidelity (Sugai et al., 2007; Sugai et al., 2010). Implementation strategies that build local capacity for these factors and practice "continuous regeneration" through assessment, planning, and adapting their plans achieve sustainable and effective SWPBIS development (Sugai et al., 2007).

Implementation Fidelity Assessment

To facilitate implementation of research-based practices, Bond, Becker, and Drake (2011) discuss the importance of specifically defining evidence-based practice program models and their component factors, and developing psychometrically valid scales or methods of evaluating compliance with those models. As discussed by Kratochwill and Shernoff (2004), implementation of evidence-based practices in schools and children's natural settings requires evaluation of the transportability of those practices (such as SWPBIS), including how contextual variables impact implementation

procedures and outcomes. Therefore, in measuring program level implementation fidelity, scales should be developed to have discriminative validity, to indicate when a program is being implemented in adherence to the research-supported model and demonstrating predictive validity with established implementation measures and desired program outcomes (Bond et al., 2011). In discussing consultation research more broadly (including from the individual level to systems level), Sheridan, Swanger-Gagné, Welch, Kwon, and Garbacz (2009) assert that the criteria for determining fidelity should include those components of the intervention that are most necessary for efficacy of that intervention and have been shown to have a research-supported link to the desired outcomes of that intervention. Assuming that the procedures developed by the researchers were followed and found to have empirically-supported beneficial outcomes to the client or system, those programs which are implemented with high levels of fidelity, programs meeting implementation fidelity standards should have improved outcomes when compared to programs implemented with lower fidelity.

Moreover, the assessment of implementation fidelity serves as a means for schools and districts to progress monitor their efforts to develop SWPBIS. Given the high cost of SWPBIS program building and implementation (Horner et al., 2012), school administrators and district supervisors have a vested interest in consistently making progress toward full implementation. Implementation fidelity measures, such as the SET and the TIC, are not only useful for assessing a school's overall level of implementation, but also for identifying areas in which the school is excelling or might need continued support or training. As discussed by Stokes and Baer (1977), generalization of skills often does not adhere to the "train and hope" model, wherein skills are taught and

teachers step back, “hoping” that the student will continue to build that skill set and naturally generalize those skills to novel situations or problems. In SWPBIS program building, initial training is not intended to be followed by a passive hope for continued development, but rather is intended to be continually addressed by the school PBIS team through support from a consultant, but also through self-assessment or outside evaluation of implementation fidelity (Sugai et al., 2010). Therefore, at the item level, measures of SWPBIS implementation fidelity can provide school-based teams with critical information about the areas in which they need continued training, external support, a new approach to problem solving, or greater resources (Sugai et al., 2011; Todd et al., 2004). To that end, the development of reports and action plans based on implementation fidelity measures should be considered as important an action as the assessments itself.

Content Validity

Validity is “the most fundamental” psychometric consideration in the development and revision of any scale or assessment tool (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11), allowing for users to interpret the results of a given measure with confidence. Though validity may be conceptualized in different ways, The Standards for Educational and Psychological Testing (AERA et al., 2014) emphasizes that validity refers to a “unitary construct” (p. 14), of which one may find “types of validity evidence, rather than distinct types of validity” (p. 11) from test content and structure, its predictive relationship with other measures or criterion tests, and consideration of group differences in its use. Therefore, the best practice according to the Standards is to integrate all evidence of an assessment’s

validity to specify the exact uses, interpretations, and conditions under which that test or measure will provide valid information about a specific construct (AERA et al., 2014, pp. 21-22). This theory of a single unified validity is careful not to refer to unique types of validity (e.g., "content validity" or "criterion validity"), arguing that different facets or evidential bases of validity must all be considered together and independently may give the appearance of sufficient validity when a test lacks overall validity for decision making. For the purposes of this study, the AERA, APA, and NCME's (2014) position on validity will be valued in regard to the importance of looking at multiple facets of validity to determine usefulness and appropriateness, however the more common terminology presented by Messick (1988) and Murphy and Davidshofer (2005) will be used to differentiate between the different types of validity evidence. Specifically, content and concurrent validity will be discussed as areas demanding unique analyses and considerations, though with the understanding that there must be a comprehensive integration and consideration of all types or aspects of validity of any given measure.

One type of validity that researchers must consider in the development and use of any rating scale, including a measure of implementation fidelity, is whether that assessment has content validity. Content validity refers to whether the behaviors or attributes being measured by the tool are representative of the construct that it is supposed to measure (Murphy & Davidshofer, 2005). In other words, content validity is the consideration of whether the items being measured truly reflect the core components of the construct such that the results of the measure will provide an accurate indication of the levels of that construct. Rather than being a static construct, Murphy and Davidshofer (2005) explain that the validity of any given test or scale should be an ongoing process

and may change along with expert opinion and experience of the construct being measured. Particularly as understandings of constructs evolve, content validity should be reconsidered and re-examined to ensure continuing validity in decision-making utility.

Assessment of SWPBIS Implementation Fidelity

Horner and colleagues (2004) cite increasing trends toward researching the effects of SWPBIS as a motivating reason for the development of tools to assess SWPBIS implementation fidelity. The SET and other tools will therefore serve as a means for researchers to determine whether a school is implementing procedures associated with SWPBIS, whether changes in student or staff behavior may be related to implementation status, and measuring the effects of SWPBIS training or coaching (Horner et al., 2004). Furthermore, implementation status has been demonstrated to have a relationship with desired and meaningful outcomes for schools and students. For example, in a validation study of the SWPBIS implementation measure, the Benchmarks of Quality (BoQ; Kincaid, Childs, & George, 2005), Cohen and colleagues (2007) found that schools meeting the 70% implementation standard demonstrated significantly greater decreases in ODRs than schools scoring below 70% on the BoQ.

Level of SWPBIS implementation fidelity has also been linked to positive improvements in school climate and teacher well-being. Ross and colleagues (2012) anonymously surveyed teachers on their well-being and compared the responses of teachers from schools that had achieved a passing score on the SET (80% or higher) with the responses of teachers from schools that had failed to pass the SET. The survey included items related to teacher stress level, burnout, sense of accomplishment, and sense of efficacy. Schools' implementation status on the SET was found to be strongly

and positively related to teachers' reports of their efficacy in the classroom and was associated with lower rates of emotional exhaustion. Teachers from schools that had achieved successful implementation status on the SET were more likely to report a higher sense of personal accomplishment. This relationship was moderated by the socioeconomic level of the school community, such that teachers in lower socioeconomic communities demonstrated a more significant improvement in well-being from successful implementation of SWPBIS. The researchers attributed the positive benefits of SWPBIS on teacher well-being to the procedures' emphasis on the use of evidence-based practices and expectations for teaching students specific behavior rules.

Tools for Evaluating SWPBIS Implementation Fidelity

Given the importance of fidelity to intervention procedures when implementing an evidence-based intervention or program, SWPBIS researchers and practitioners have developed a number of different tools or measures for evaluating SWPBIS implementation fidelity. These assessments vary in their intended user, frequency of use, as well as in the specific components of SWPBIS that are addressed. At the first tier of SWPBIS, implementation fidelity rating scales, surveys, and checklists allow PBS teams, administrators, or outside evaluators to assess whether a school has developed the core components of SWPBIS. In addition to providing an overall indication of implementation success, tools such as the SET, BoQ, or TIC also provide a means of making decisions regarding school and student needs as well as prioritization of resource allocation for SWPBIS development (Solomon et al., 2012). When conducted regularly and systematically, implementation fidelity assessments allow for the development of action plans which allow for the tailoring of SWPBIS components to the precise needs

and context of a school (Sugai et al., 2007). Though schools often look to find that they have passed the standard for implementation success, at an item level, these assessments are designed to provide schools with specific feedback regarding the SWPBIS components that are still in need of development.

School-wide Evaluation Tool

The School-wide Evaluation Tool (SET) was developed by Sugai and colleagues in 2001 with the intention of being an assessment tool that researchers could use to evaluate SWPBIS implementation and effectiveness (Horner et al., 2004). It is also cited as being the first validated SWPBIS implementation measure (Cohen et al., 2007). Horner and colleagues (2004) describe the seven areas of SWPBIS assessment included on the SET as originating from seven key practices of SWPBIS. The SET was developed to include 28 items, each of which fit within one of the seven categories of SWPBIS development and practice. These categories represent the factors that researchers have determined necessary for successful implementation of SWPBIS procedures with fidelity (Lewis & Sugai, 1999). The seven categories include: Expectations Defined; Behavioral Expectations Taught; On-going System for Rewarding Behavioral Expectations; System for Responding to Behavioral Violations; Monitoring and Decision-Making; Management; and District-Level Support (Sugai et al., 2005).

One way in which the SET differs from other SWPBIS implementation fidelity measures is that it was designed to be used by researchers or agencies outside of the school being assessed (Horner et al., 2004). Whereas the TIC, BoQ, and EBS Survey were developed with the intention of being completed by the school's own SWPBIS team or through collection of faculty members' perceptions, the SET's administration

procedure allows it to be used by someone unfamiliar with a given school's specific SWPBIS plan. The SET begins with an administrator interview and a review of SWPBIS-related documents that allows the evaluator to familiarize him/herself with the plan before beginning direct observation of the school and interviewing of students and faculty (Todd et al., 2004). For this reason, it is easily utilized by state agencies performing annual implementation assessments (as is done in many states such as Pennsylvania and Maryland) or by researchers looking at fidelity of SWPBIS to establish relationships with outcome measures. The authors of the SET also developed a training manual (Todd et al., 2004) that provides a standardized method of introducing and teaching researchers, practitioners, and graduate students to administer the SET with inter-observer agreement of 99% (Horner et al., 2004).

Though the SET has the benefit of allowing "blind" unbiased assessment of a school's level of SWPBIS for research or policy purposes, it requires that an individual outside the school be hired or contracted to conduct the assessment, therefore requiring greater allocation of funding and time than those assessments conducted by school-based teams. Therefore it is critical that the scores yielded by the SET provide a valid and reliable evaluation of a school's level of implementation. While research on the SET has established that it has acceptable levels of reliability (Horner et al., 2004; Vincent, Spaulding, & Tobin, 2010), as discussed below, there remains a lack of research regarding the validity of the SET's construction, weighting, and scoring.

Establishing the Validity of the SET

To validate the SET, Horner and colleagues gathered SET data from 45 schools. Data were collected by evaluators and observers who had been trained using materials

and standardized procedures from the SET implementation manual (Todd, Lewis-Palmer, Horner, Sugai, & Phillips, 2002), and inter-observer agreement from school-based SETs was found to be 99%. Internal consistency was evaluated both in regard to item/subscale score consistency as well as item/total score consistency and found to be at acceptable levels with an overall alpha of 0.96. Subscales demonstrated suitable internal consistency with the overall score as well, ranging from 0.69 ("District Support") to 0.89 ("Behavioral Expectations Taught"). A single item (#13 pertaining to the school's crisis plan) was found to have questionable consistency with the other items on the SET, but was not dropped or removed from the SET scoring. Test-retest reliability was conducted across eight schools, finding that scores were consistent across a period of two to three weeks (97.3% consistent). Next, the SET's construct validity was assessed by establishing convergent validity between the SET's total score and a score based on 15 items from the EBS Survey (Sugai, Todd, et al., 2000; referred to in the study as the *Effective Behavior Support: Self-Assessment Survey*). Because the EBS Survey is a broad measure of implementation fidelity that assesses not only Tier I SWPBIS implementation, but also nonclassroom settings and includes family considerations and perspectives, only a portion of its items were used in the validity analyses. Comparing only those EBS Survey items pertaining to Tier I, the SET demonstrated a high Pearson correlation, $r = .75$ ($p \leq 0.01$). A preliminary assessment of internal validity indicated that subscale intercorrelations ranged from r = .44 to r = .81, establishing "sufficient empirical association to be interpreted as components of the SET total score" (p. 8). One criticism of this study, however, is that while the SET items are scored on an ordinal scale (indicating different non-interval levels of implementation), they were treated as

through they were on an interval scale and used to find mean scores as well as in Pearson correlations and other analyses. Though problematic in relation to Stevens' hierarchy of measurements (1946), the use of ordinal scale values in finding mean scores and correlations is common practice in SWPBIS research, as well as more generally across applied science fields that utilize rating scales (Choi, Peters, & Mueller, 2010).

While Horner and colleagues (2004) found the SET to have adequate test-retest reliability and inter-observer agreement, moderate to moderately high scale intercorrelations, and to be sensitive to changes before and after SWPBIS training and implementation, the SET lacks direct observation or measure of changes in student academic or social behaviors. The authors offer that future research should investigate the relationship between SET scores and school measures of ODRs, attendance, special education referrals, as well as student academic performance (Horner et al., 2004); however, no known research has yet directly looked at a way to incorporate these or other direct outcome measures of school discipline and performance into the SET.

There are two primary methods of scoring the SET. Based upon Horner and colleagues' preliminary validation of the SET (2004), it was suggested that the standard for successful SWPBIS implementation be set at earning both an 80% overall mean score across the seven categories of implementation factors as well as an 80% in the area of "Behavioral Expectations Taught." This recommendation was based on Horner and colleagues' (2004) clinical observations that in addition to having an overall implementation score of 80%, "change in student behavior is unlikely before a school *teaches* the school-wide expectations" (p. 11). This 80/80 standard is applied in clinical uses of the SET to evaluate SWPBIS implementation by agencies and schools. At times,

however, researchers (e.g., Cohen et al., 2007; Horner et al., 2009) use the total score instead of the 80/80 standard. This is likely explained as a decision by the researchers to utilize the total implementation for more straightforward statistical analyses, as use of the overall SET score allows for more simple analyses based on a single continuous variable, rather than the creation a dichotomous variable to account for meeting or failing to meet the 80/80 standard. The continuous variable also allows for greater sensitivity when measuring implementation status and growth over time.

Following its initial publication (Sugai, Lewis-Palmer, et al., 2001), validation (Horner et al., 2004), and years of use, Vincent and colleagues (2010) reassessed the validity of the SET using a significantly larger data set collected through PBSsurveys.org (now PBISapps.org), a systematic data collection system for SWPBIS implementation tools run by Educational and Community Supports at the University of Oregon. Using a data set of 1352 schools from preschool through high school, 65% of which had earned passing scores (80/80) on the SET, Vincent et al. (2010) looked more specifically at differences in the use of the SET across school levels. Of relevance to the current study, the researchers obtained a sample of 833 elementary schools that had entered data for the SET. At the elementary level, the SET had adequate overall internal consistency (r = .850), though additional consideration of confidence intervals indicated that the "Teaching Behavioral Expectations" and "Management" subscales may be less well defined and contain more error than the other subscales. This finding is concerning given the relative importance given to "Teaching Behavioral Expectations" within the SET's 80/80 scoring standard. This study also evaluated the concurrent validity of the SET with the TIC; however, the analyses were based on a version of the TIC which has since

been significantly revised and the researchers included only small subset of the total number of items on the TIC.

Beyond studies specifically validating and establishing the reliability of the SET, the validity of the SET has also been studied in regard to its concurrent validity with other measures of SWPBIS implementation fidelity. Cohen and colleagues (2007) used the SET as a comparison measure during the development of the BoQ. Though the two measures assess many of the same components of SWPBIS, the BoQ goes into more detail across 53 items (compared to the 28 items of the SET). To that end, though on average BoQ scores were between 9 and 15 points lower than SET scores, the total BoQ score and total SET score showed a moderate correlation of 0.51 ($p < 0.05$).

The SET stands apart from other measures of SWPBIS fidelity in that it was developed with the intention of being used by agencies outside of schools, including state departments of education, educational service agencies, and researchers. For example, the SET is used by publically funded PBIS agencies in states such as Florida (Kincaid, Childs, Blase, & Wallace, 2007) and Maryland (Barrett, Bradshaw, & Lewis-Palmer, 2008), and researchers have used the SET in SWPBIS studies across the United States (e.g., Horner et al., 2009). In contrast, the BoQ, TIC, and EBS Survey all are completed either by the school-based PBIS team or through surveys completed by the school staff. The uniqueness of the SET in its ability to be administered by individuals outside of the school makes it critical that the field continue to validate its content and scoring methods.

Another unique feature of the SET compared to other SWPBIS implementation measures is that it directly addresses whether the school and district have allocated specific funds in their annual budget for "building and maintaining" a SWPBIS system.

Though the steps involved in implementing SWPBIS do not formally include establishing a funding source, garnering the financial support of the district, school board, and school community through a SWPBIS budget line is important for the costs incurred by components of an efficacious multi-tiered school behavior plan. Horner and colleagues (2005) specify that investment in SWPBIS development requires explicit funding for the development of a PBIS team, training, staffing (to allow staff members time to work on the school-wide plan), and tangible reinforcers or rewards for reinforcing appropriate student behaviors. Additionally, schools may need to invest in the development or purchase of data collection systems, such as SWIS (May et al., 2006). Though the TIC asks about the development and presence of SWPBIS components that cost money, it does not address whether the school has a specific plan for funding critical aspects of their program. When undertaking the development of SWPBIS, Horner and colleagues (2012) suggest that in addition to formally emphasizing behavior concerns into the district's goals, the school must hire or contract with training resources to build capacity, work with national consultants, develop data collection systems for both student outcomes and implementation fidelity, as well as plan for the reorganization of staff for training and team meetings. Therefore, to begin a new SWPBIS initiative across multiple schools, it is estimated to cost between $5000 and $10,000 per school for the first two years of implementation (Horner et al., 2012). Based on the resources required to begin and develop sustainable implementation, it is critical that the validity of the SET be reconsidered to ensure that it is valid tool for the assessment of areas of continuing need for SWPBIS development as well as overall implementation.

Validity Considerations in Using the SET

In their first article documenting the psychometric properties of the SET, Horner and colleagues (2004) refer to Messick's (1988) theory of validity as the basis for the conceptual logic with which the SET was developed and initially evaluated for its psychometric characteristics. Messick (1988) proposes that the "key validity issues are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use" (p. 33). He emphasizes the importance of the meaning and usefulness of scores in relation to the construct they are trying to measure as well as the relatedness of the construct and content that is being assessed. When establishing content validity, Messick (1988) explains, "content-related evidence usually takes the form of consensual informed judgments about the representative coverage of content in a test and about its relevance to a particular behavioral domain of interest" (p. 38). Once content validity is established, however, the test must still be evaluated with regard to the usefulness of the results of that test and the validity of the inferences that can be drawn from those results (Messick, 1988).

The SET is not without its critics in the field of SWPBIS. In particular, Cohen et al. (2007) critiqued the SET in their study to validate the BoQ as a measure of implementation fidelity. Though they acknowledge that the SET is both psychometrically sound and widely used, the authors are critical of the measure's time intensive nature (requiring two to as many as six hours per administration depending on the size of school and experience of evaluator) and disruption of student and staff routines for interviews. Most critical to the discussion of the SET's validity for

implementation evaluation and decision making, Cohen and colleagues (2007) observe that "schools can score over 80% on the SET without having many of the critical features of SWPBS, such as lesson plans and an evaluation plan, in place" (p. 204).

This final critique resonates with the lack of clarity regarding the weighting and scoring procedures of the SET. For the overall score of the SET, though each of the seven categories on the SET is weighted as one-seventh of the total score, the number of items in each category ranges from two items in School District Support and Behavioral Expectations Defined to as many as eight items in Management. Therefore, there is a large discrepancy in how much weight a single item carries in "School District Support" ($1/14^{th}$ of the overall score) compared to a single item in "Management" ($1/56^{th}$ of the overall score). Though the grouping of items by category is useful for completion of the SET, the manner in which they are scored imposes an artificial weighting on their relative importance. Therefore, though the items on the SET may be drawn from theoretical bases (e.g., Lewis & Sugai, 1999), that does not necessarily indicate that the way in which the items are weighted or scored will translate into a valid indication of whether a school has implemented the most critical aspects of SWPBIS.

It appears that this inconsistency in weighting was apparent at the time of the SET's validation, as Horner and colleagues (2004) added the secondary implementation criteria of requiring 80% fidelity in the "Behavior Expectations Taught" category. While their justification for this scoring is logical (children must be taught expectations before they can reasonably be expected to demonstrate those behaviors), the 80/80 scoring standard still leaves many critical components of SWPBIS implementation as counting for as little as 1/56 of the overall score. This scoring method is called further into

question by Vincent and colleagues' (2010) re-examination of the SET, in which they found that the "Behavioral Expectations Taught" category had the second widest confidence interval of the seven categories, which they posit may indicate that the category contains significant amounts of error and may be "less well defined than the remaining key features of SWPBIS" (p. 171). Based on this finding, perhaps the "Behavioral Expectations Taught" subscale is too unreliable of a measurement on which to judge teaching of behavioral expectations or to make decisions based on the 80/80 standard. In the current study, the relative weightings of each item on the SET will be considered independently by experts in order to establish whether all of the items are critical to SWPBIS implementation and to determined how to most fairly evaluate whether a school is implementing those components of SWPBIS that are most important for durable and sustainable system change.

With consideration of the *Standards for Educational and Psychological Testing* (2014*)* developed by the AERA, APA, and NCME, the SET also would benefit from re-examination and re-validation. For example, Standard 1.11 states:

> When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified. (p. 26)

Though Horner and colleagues (2004) explained the basis for the categories included on the SET, the scoring and weighting of specific items are not justified with theory or quantitative evidence, nor as the *Standards for Educational and Psychological Testing* (AERA et al., 2014) continue, have the SET developers "illustrat[ed] the relevance of

each item and the adequacy with which the set of items represents the content domain" (p. 26). Though the SET items represent a reasonably comprehensive snapshot of components of SWPBIS plans, consideration must be given to which components might be under-or over-represented by its current implicit weighting and scoring rules.

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), validation of any measure is the responsibility both of the test developer and the test's users, providing support for the construct being measured, how it is represented by the items in the assessment tool, and evaluating the appropriateness of its use within context (p. 13). Therefore, as the SET has been used, SWPBIS implementation procedures have been refined, and other implementation fidelity tools have been developed, it is important for experts in the field to examine its continuing content validity and the usefulness of the information it provides. Furthermore, as researchers and practitioners have expanded their work with SWPBIS and it has extended across a greater range of states, school settings (e.g., urban vs. rural), increasingly diverse student populations, they may have gained invaluable clinical and quantitative knowledge of which aspects of SWPBIS seem to have the greatest impact on successful or sustainable implementation.

Though best practices in individual, classroom, and school-wide consultation include measurement of implementation fidelity (Erchul & Martens, 2002), there remains a lack of consensus regarding the development of fidelity assessment tools, how to determine which components are most critical to measure, and the psychometric value of such tools (Sheridan et al., 2009). Sheridan and colleagues (2009) assert that current practices in developing fidelity measures would benefit from greater "clarity" in the

process of developing fidelity measures. Specifically, they pinpoint nine areas of vague standards or lack of consistent consideration in fidelity measurement development, including aspects such as the selection of factors that most directly predict intervention outcomes, the "match of measurement tool to intervention components" (p. 479), reliability across various sources of assessment information, and how easily the assessment of fidelity might be completed. Most germane to the discussion of the content validity of the SET, the authors specifically note a lack of clarity around procedures pertaining to "the relative weighting of specific plan components," "the appropriate metric that relates empirically to outcomes (e.g., item level, global score, or consistency in implementation over time)," and "the sensitivity of measures to assess meaningful treatment components" (i.e., is the measure assessing those implementation factors which are most critical) (p. 479). These concerns parallel the psychometric and design elements of the SET that are lacking in justification and empirical support. Therefore, the current study aimed to look at the relative importance and weighting of the implementation factors included on the SET, specifically evaluating the SET's content and scoring based on the opinions of experts in the field of SWPBIS.

Item Weighting of Rating Scales

When an assessment is in the process of being created, its developers must not only determine what content areas and components should be included and how they should be measured, but also must justify how those components are weighted to obtain a valid measure of the targeted construct (AERA et al., 2014). As discussed above, the SET was created based on seven areas of SWPBIS implementation considered critical for successful implementation (Horner et al., 2004; Lewis & Sugai, 1999). Though the items

and grouping of items into categories were based in theory, the weighting of items was determined somewhat arbitrarily based on the number of items per category and the overall mean across categories. According to the literature regarding test development and item weighting, often the most effective and efficient form of weighting is assuming equal, or natural, weighting of all items (Wang & Stanley, 1970). Therefore, to justify the effort of differentially weighting items, test developers must demonstrate that item weighting provides a benefit over equal weighting with regard to test validity and/or reliability.

Despite arguments for natural or equal weighting of test items (Wang & Stanley, 1970), researchers continue to develop methods of differentially weighting items and subscales within a larger assessment. This is ostensibly because, despite support for unit weighting, logically it makes sense that more important or foundational items/skills should be given greater weight in determining overall scores (Wang & Stanley, 1970). For example, Feldt (2004) discussed the issue of weighting algorithms within the context of teachers' algorithms for determining test scores or overall course grades. Specifically, he discusses how in constructing tests, it is not always feasible for teachers to include sufficient numbers of items related to critical higher-order thinking skills (due to their difficulty or the time necessary to address such tasks), and therefore these items may need to be weighted more heavily than items addressing more specific facts or instruction-based items. While less formal than assessment development within a research-context, similar logic is often used when researchers approach test development, as it is natural to have a priori judgments or expertise that indicate all factors may not contribute equally to a desired outcome.

Even when addressed from a more formal and standardized testing perspective, differential item or component weighting is generally based on a series of arbitrary steps, decisions, or judgments. Arthur, Doverspike, and Barrett (1996) present a weighting procedure they call Relative Content Contribution (RCC), which allows component weights of an assessment to be determined based on content-related judgments, laying out a system of collecting data on relative importance and assigning weights to measure work behaviors of employees. Despite their attempts to objectively determine weightings, the authors concede that the process is somewhat arbitrary and subjective in how the subjective ratings of behaviors are converted into weightings. Feldt (2004) also discusses two types of scoring methods for establishing differential item weightings: weight the total scores across each area differently or convert raw scores into standard scores which are in turn multiplied by constants to reflect their relative weightings. Each of these, however, he suggests are typically done in an arbitrary manner, based on the best estimates of the teacher or individual designing a test. In cases of standardized or widely used tests, however, it is important that test component ratings are determined in a more deliberate and validated manner, establishing a reasonable argument for why some items or groups of items are weighted more heavily than others.

Wang and Stanley (1970) reviewed the empirical support for various methods of differential weighting at both the item level and response category level within tests. In their summary of research on differential weighting in education, they identify a number of approaches to weighting subscales and items within assessments, including natural (or equal) weighting, a priori weighting based on judged importance of items, multiple regression to weigh predictor variables more heavily (but requires a criterion measure),

and even more complex methods designed to ensure equivalent correlations within composites or equivalent weighting toward the assessment's total variance (Wang & Stanley, 1970). Despite the variety of weighting methods and the commonality of weighting in both informal and formal testing, Wang and Stanley concluded, "weighting of test items was shown repeatedly to be ineffective, or so slightly effective as to be impractical" (p. 699). Though they discuss that item weighting has logical appeal (p. 663), the efforts put in by researchers to weight test items or test composites accordingly have been met with limited success with regard to improving validity outcomes.

In considering the current scoring process of the SET, Horner and colleagues (2004) created between two and eight items across seven categories of SWPBIS implementation factors. By structuring the SET so that variable numbers of items were grouped within equally weighted subscales, the SET's design impacts the weight each item contributes to an overall implementation score. Feldt (2004) draws a contrast between how component weighting can be accomplished either *directly* by multiplying raw item scores by their relative weightings or *indirectly* through the scoring methods and values assigned by the test developer. He uses the terms "effective length" or "functional length" (p. 189), which refer to the idea that test items or sections that have a greater number of possible points have a greater "implicit weight" than those with fewer possible points. Within the SET, implicit scoring weights are created by differences in the number of items in each subscale (e.g., eight items in the "Behavioral Expectations Taught" subscale compared to two items in the "District-Level Support" subscale) and then by averaging all seven subscales to give them equal weight. The process of finding a mean score for each subscale before averaging the seven categories or subscales together,

however, indirectly assigns a weight to each item based on its category. Therefore, based on the recommendations of Wang and Stanley's review (1970), it would be worthwhile to examine whether a SET with equally weighted items might be as effective or more effective at measuring SWPBIS implementation fidelity than its current method.

Current Study

The current study re-examines the validity of the SET to establish whether an alternative approach to weighting and scoring the SET might provide a more valid assessment of a school's level of SWPBIS implementation fidelity. The first study discussed in this paper evaluated the content validity of the SET by surveying experts in the field of SWPBIS to obtain their ratings of the relative importance of each item on the SET. This content validity study was conducted to provide an indication of how well the SET matches the content and structure of SWPBIS implementation factors (Murphy & Davidshofer, 2005). The experts' ratings were obtained to allow for consideration of which SET items are most and least critical to sustainable SWPBIS implementation. Though natural weighting of items is most often recommended in test development (Wang & Stanley, 1970), the mean expert ratings were used to develop a new system of item weighting for scoring the SET using the a priori method, in which "nominal weights are assigned on the basis of judgments or ratings" (Wang & Stanley, 1970; p. 668). The goal of this first study was to re-evaluate the current scoring method of the SET, and consider whether an alternative method might give more accurate weighting to those items that are most critical and pivotal to successful and sustainable SWPBIS implementation, while minimizing the impact of or excluding those items judged to be least critical.

Based on the content validity study, the second study more closely examined whether the information gained from the expert ratings might be used to improve the concurrent validity of the SET. To that end, the two original scoring methods of the SET (as established by Horner et al. [2004]) as well as an unweighted version and a re-weighted version were compared to a second well-established tool for assessing SWPBIS implementation fidelity, the TIC (Version 3.1). Furthermore, analyses on the two novel SET scoring approaches were repeated with the exclusion of those SET items that fell in the bottom quartile of the experts' mean ratings. This study used existing data collected by the Educational and Community Supports program at the University of Oregon to compare the concurrent validity of a total of four novel SET scoring approaches (unweighted; reweighted; unweighted with dropped items; reweighted with dropped items) and the two established SET scoring methods with the TIC.

CHAPTER 3

METHODOLOGY

This dissertation research included two separate studies of the SET, looking at its content and concurrent validity, respectively. The first study addressed the content validity of the SET with regard to the alignment of its content and structure through the examination of experts' opinions of the relative importance of items on the SET. Based on the mean ratings of importance obtained from this survey, four novel scoring approaches were developed, including: unweighting SET items, reweighting the SET items according to the expert ratings, and removing items rated within the lowest quartile from both the unweighted and reweighted methods.

Study 1

Participants

For Study 1, experts in the field of SWPBIS were recruited to participate in an online survey regarding the relative importance of items on the SET to sustainable SWPBIS implementation. For the purposes of this study, it was determined that a group of approximately 45-60 researchers would be initially recruited based upon the number of articles they have published on SWPBIS. This number of participants was selected based on guidelines provided in the literature on seeking expert consensus, indicating that a panel of 15 to 30 individuals is necessary when the group is composed of experts within the same field (Clayton, 1997). Therefore, a goal was established to find 45-60 experts, with the expectation that a return rate of 30% or more would yield a panel of the

recommended size. It was anticipated that this return rate was attainable due to the low demands on participant time and relevance of this project to the SWPBIS researchers. Because the participation of the experts would take place over email and through an internet-based survey engine, participants were not asked or required to travel or directly interact with the researcher.

Search Criteria

To find participants for Study 1, researchers in SWPBIS were identified using an electronic literature database search via the PsycINFO and Education Source databases (EBSCO Publishing, 2014). The search was initially extended back to 1999, when the first articles were published specifically about SWPBIS systems; however, this search was extended beyond this date range to ensure inclusivity. These databases were searched using the terms "schoolwide positive behavior support," "school-wide positive behavior support," "SWPBIS," "SWPBS," and "SW-PBIS" to account for variability in the acronym for SWPBIS over time. The results of these searches were cross-referenced with a search specifically within journals identified within Chitiyo and colleagues' (2012) meta-analysis as publishing articles on SWPBIS research. The journals searched included: *Psychology in the Schools, Journal of Positive Behavior Interventions*, *Journal of Emotional and Behavioral Disorders,* and *Education and Treatment of Children*. Additionally, other journals identified through the present study's literature search (specifically, *School Psychology Review, School Psychology Quarterly,* and the *Journal of School Psychology*) were searched with the same keywords. Finally, the list was cross-referenced with a list of SWPBIS research and resources provided on OSEP's Technical Assistance Center on PBIS website (2009a) and articles included in recent

meta-analyses on SWPBIS (Chitiyo et al., 2012; Solomon et al., 2012). In total, these searches yielded 229 articles focusing on SWPBIS and school-wide level behavioral programs.

Expert Inclusion Criteria

Based on the search parameters described above, the authors of the 229 articles were catalogued and sorted according to number of publications. Sixty-four authors were identified as having published three or more articles on SWPBIS. Given that this met the search goal of 45-60 researchers to recruit, it was not necessary to include researchers with fewer than three publications. Of the 64 experts identified, contact information was obtained for 58 researchers. Within this group, the range of publications was 3 to 36, with the median number of publications being four articles and an interquartile range of three to six articles. By identifying experts through a review of publications, this group likely included primarily doctoral level researchers, along with some masters level SWPBIS researchers or practitioners who work within schools and at educational agencies. All invited participants were adults over the age of 18.

Participation was not limited to individuals with doctoral degrees in school psychology or related fields, with the intention that practitioners or graduate students who have gained significant experience in the field and expertise through research be included as well. Additionally, inclusion was not deliberately limited to SWPBIS experts within the United States; however, the literature search was limited to English-based publications and the research survey and contacts were completed in English. Experts were contacted directly regarding participation via emails obtained either through their publications or their academic institutions. A copy of the email invitation sent to the 58

experts and the informed consent form are included in Appendices B and C, respectively. The email included information regarding the purpose and nature of the study, the institutional review board's approval and contact information, and a link to the survey.

Measures

School-wide Evaluation Tool (SET), Version 2.1

The SET, currently in Version 2.1 (Sugai et al., 2005), is one tool for assessing SWPBIS implementation fidelity. Developed in 2001 (Sugai, Lewis-Palmer, et al.), the SET was intended to be a measure that could be used by assessors outside of the schools such as researchers and educational/state agencies (Horner et al., 2004), and the SET may be administered in approximately two hours. The SET is designed to capture the context-specific aspects of a school's SWPBIS plan, including the school's rules, method of reinforcement, teaching of the plan to staff and students, as well as the school's adoption and use of data collection. These aspects of implementation are assessed through an interview with the school's administrator, a review of SWPBIS-related permanent products, observation of SWPBIS indicators throughout the school building, as well as interviews with both students and staff. To that end, the SET includes 28 items across seven categories of implementation components including: Expectations Defined; Behavioral Expectations Taught; On-going System for Rewarding Behavioral Expectations; System for Responding to Behavioral Violations; Monitoring and Decision-Making; Management; and District-Level Support (Sugai et al., 2005).

Initial validation of the SET based on data collected from 45 schools implementing SWPBIS indicated that the SET had psychometrically strong levels of internal consistency at the item/subscale and item/overall score level, with an overall

alpha of 0.96 (Horner et al., 2004). Additionally, Horner and colleagues (2004) found high levels of inter-observer agreement (97%) among individuals who had been trained using the SET manual, strong test-retest reliability over a span of two to three weeks, as well as strong concurrent validity with Tier I items from the EBS Survey ($r = .75$; $p \leq 0.01$). Preliminary subscale intercorrelations were conducted to obtain a measure of internal validity, ranging from r = .44 to r = .81. After several years of data collection through Educational and Community Supports and PBISapps.org, Vincent and colleagues (2010) conducted a re-examination of the SET's validity on a larger sample of schools' data. Using a sample of 833 elementary schools with SET data, the researchers found that the SET demonstrated adequate overall internal consistency ($r = .850$). Along with this finding, the authors also found that the confidence intervals for the "Teaching Behavioral Expectations" and "Management" subscales were larger than the other subscales, indicating that they may be less well defined and contain more error.

As discussed previously, there are two methods for scoring the SET, a dual 80% overall and 80% in "Behavioral Expectations Taught" (Horner et al., 2004) and obtaining an overall average score by averaging ratings in each of the seven areas and then averaging those scores (used in research studies such as Horner et al., 2009). The exact order and wording of items from the SET were used in this study to obtain expert ratings of their relative importance for sustainable SWPBIS implementation.

SET Validation Survey

For the purposes of this study, an online survey was developed to solicit ratings of the relative importance to SWPBIS implementation for each of the 28 items on the SET. Given that this study seeks to re-examine the content validity of the SET, it was

determined to be beneficial to disclose the goal of re-weighting the SET by determining the relative importance of each item, and therefore the items are presented with the exact wording of the SET, in the same order they appear on the rating form. The items were presented together on the same survey page and participating experts were asked to rate the items on a Likert-type scale from 1 ("Not at All Important for Sustainable Implementation of SWPBIS") to 5 ("Critical for Sustainable Implementation of SWPBIS") (See Table 1). The presentation of all SET items at once was selected so that participants would be able to consider relative importance of each item in relation to other items on the SET and adjust their ratings if desired. After rating each SET item, the participants were also asked whether they would add any implementation components to the SET as well as what SWPBIS implementation tool they prefer to use and why. In addition, the survey asked participants to respond to items regarding demographic information, such as highest level of education, occupation, years of experience with SWPBIS, ethnicity, and race. The survey was administered through an online survey service, www.surveygizmo.com. See Appendix D for a copy of the survey description and items as presented to participants.

Table 1. SET Validation Survey Rating Scale Options

Scale Option	Description of Rating
1	Not at All Important for Sustainable Implementation of SWPBIS
2	Helpful but not Necessary for Sustainable Implementation of SWPBIS
3	Important, but not Necessary for Sustainable Implementation of SWPBIS
4	Very Important, but not Critical for Sustainable Implementation of SWPBIS
5	Critical for Sustainable Implementation of SWPBIS

Design and Procedure

Experts identified as being the most prolific researchers in the field of SWPBIS (defined as having three or more published articles on SWPBIS) were emailed directly by the researcher to participate in the online SET Validation Survey. Initial contact was made with the potential participants individually over email, and included information regarding how and why the individual was selected for participation and a discussion of the importance and benefits of the research study. Participants were invited to learn more about the study and their participation by clicking a link to the survey at SurveyGizmo (2014; www.surveygizmo.com), which was provided in the email. The survey was described as a brief online measure, taking approximately twenty minutes to complete. By clicking the link in the invitation email, prospective participants were presented with a description of the research, a rationale for the study, and an explanation that the goal of the study was to examine the content validity of the SET by looking at the relative importance of each item to SWPBIS. Additionally, this page presented information regarding the risks and benefits of participation, institutional review board approval, and contact information. Experts who consented to participate marked their assent and were then directed to the online survey, while anyone who refused consent was exited from the website. The use of an internet-based survey was selected for ease of responding, the convenience of participants, and simplicity of data-collection methods. Dillman, Smyth, and Christian (2009) suggest that a period of one week is a reasonable timeframe for participants to complete a survey; therefore, participants were given a week, with two additional reminders to participate before the survey was closed.

Efforts were made throughout the planning and design of the online survey to maximize participation and minimize unforeseen complications. Dillman and colleagues (2009) identify four factors that account for most survey difficulties and failures: coverage, sampling, nonresponse, and measurement error. Specifically, they discuss measurement error related to survey wording, participant response errors, and complications due to the formatting or medium of a survey. To avoid measurement error, efforts were made to keep the item words and response demands as simple as possible. Participants were asked to rate the relative importance of each item of the SET on a one to five Likert-type scale (Likert, 1932), with limited additional demographic questions or requests for free responses.

With regard to issues of coverage, sampling, and nonresponse (Dillman et al., 2009), they were not expected to be problematic in this study, as a specific group of survey experts in the field of SWPBIS was identified through a rigorous publication search process. Issues of bias have not been found to confound studies based on expert opinion (Hsu & Sandford, 2007), so basing inclusion on expertise was considered more important than seeking diversity across participants' backgrounds, affiliations, and collaborations. In the current study, since opinions were sought on the relative importance of items on a previously developed and validated measure (rather than trying to identify a consensus around a novel construct), the development of a diverse group of individuals from varied specialty areas was unnecessary. Furthermore, experts who participated in the development and validation of the SET were not excluded, as their judgments related to critical SWPBIS components were deemed important to the study's outcomes. It was anticipated that there would not be an excessively high rate of

nonresponse as the invited participants represent individuals who are invested in continued research into SWPBIS and its implementation. Dillman and colleagues (2009) suggest that providing information about the survey, asking the participants for their help, showing positive regard for the participants, and emphasizing the importance and context of the survey may help increase positive perceptions of the survey and likelihood of completion. In this study, no monetary or tangible incentive was offered for participation.

Data Analyses

The goal of the first study was to evaluate the content validity of the SET through re-weighting of the SET items to reflect the expert panel's consensus on the relative importance of each item to sustainable SWPBIS implementation. To address the first research question regarding the relative importance of items on the SET, the experts' individual ratings of each item of the SET (based on their online survey responses) were collected and analyzed to find the mean ratings, as well as the median, mode, and range of ratings for each SET item. As is often done in social science research (Choi et al., 2010; Norman, 2010), the ordinal Likert-type ratings of the items on the SET Validation Survey were treated as though they had interval properties so that they could be entered into calculations of mean scores. Though the use of ordinal (as well as nominal) variables in descriptive and correlational statistical analyses is problematic based on Stevens' hierarchy of measurements (1946), statisticians (e.g., Lord, 1953) have suggested that nominal and ordinal data may be used in such analyses since "…the numbers don't remember where they came from." In this particular case, it was necessary to calculate descriptive statistics for the Likert-type items to evaluate the

importance of each SET item. These statistics include mean rating, median and mode ratings, and the range of ratings. Additionally, frequency distributions were calculated for each SET item's ratings. The range of ratings and frequencies of each rating per item were examined to look for items on which the experts demonstrated an obvious split in their valuation of its importance, defined as polarized response patterns with a lack of mid-range ratings.

Development of Novel SET Scoring Approaches

The mean ratings of the relative importance of each item were then utilized to develop novel methods of scoring the SET. Based on the opinions of the SWPBIS experts surveyed, four new approaches to scoring the SET were developed to be compared to the established SET scoring methods in Study 2. As discussed by Wang and Stanley (1970), the literature on differential item weighting suggests that though there is logical appeal in weighing scale items according to their importance, such weighting methods are rarely worth the effort in development and scoring. Furthermore, the effort of weighting is rarely associated with a statistically significant improvement in validity or reliability (Wang & Stanley, 1970). This, however, has not stopped test developers from attempting to weight items or subscales. One such method is the a priori method, in which "nominal weights are assigned on the basis of judgments or ratings" (Wang & Stanley, 1970; p. 668). Given the recommendation to use equal weighting as well as the tendency for test developers to attempt weighting anyway, both methods were applied in this study. Therefore, the expert ratings were used to develop a re-weighted version of SET scoring such that each item will be weighted proportionately to its mean importance rating to contribute to an overall implementation score, while a second method removed

all weightings from the SET items, allowing for equal weights. Additionally, the experts' ratings were considered in determining which items' mean ratings were in the bottom quartile. Those items falling in the lowest quartile were excluded from the unweighted and re-weighted approaches. In sum, four new scoring methods for the SET were developed:

- Unweighted (All 28 original items having equal weight.)
- Reweighted (All 28 original items weighted according to mean expert ratings on the SET Validation Survey.)
- Unweighted with dropped items (Items equally weighted, with bottom quartile items dropped.)
- Reweighted with dropped items (Items weighted according to mean expert ratings, with bottom quartile items dropped.)

Study 2

Study 2 compared the concurrent validity of the four novel SET scoring approaches with the TIC (Version 3.1), a second measure of SWPBIS implementation fidelity, as well as the two established scoring methods of the SET (Horner et al., 2004).

Database

The second part of this study utilized an existing data set collected through the web-based application www.PBISapps.org (formerly known as psbsurveys.org), run by Educational and Community Supports at the University of Oregon. This website provides a secure and standardized system for collection of implementation fidelity data (Vincent et al., 2010), including the SET, TIC, BoQ, EBS Survey, and more targeted

fidelity tools for SWPBIS Tier II and Tier III, as well as parallel systems for the collection of ODR data, Tier II intervention, Tier III intervention, and PBIS evaluation. This allows schools and districts to track their progress and performance over time as well as create reports to guide team data-based decision making regarding training, professional development, and their SWPBIS plan. Additionally, schools using PBISapps may agree to share their school's de-identified data for research purposes. The system is designed such that schools or consultants conduct their SWPBIS implementation assessments and then local PBIS coaches or district SWPBIS coordinators can access the web-service to enter and track the assessment data. Access to www.PBISapps.org is obtained through SWPBIS coaches and leadership teams trained by the National Technical Assistance Center on PBIS (Vincent et al., 2010). Therefore, the collection of data is overseen by individuals trained in accordance with the *SWPBS Implementation Blueprint* (Sugai et al., 2010) and the *Evaluation Blueprint for School-Wide Positive Behavior Support* (Algozzine et al., 2010).

Participants

The dataset for this study was provided by the Educational and Community Supports (ECS) research program at the University of Oregon, with support from Kent McIntosh, Ph.D., Robert Horner, Ph.D., and Robert Hoselton. It was generated on March 2nd, 2014 and included data from the PBISapps.org web application for the most recent single academic year for which NCES data were available at that time, spanning fall 2011 through spring 2012. Data were compiled for this study from PBISapps, SWIS, and NCES. The criteria for inclusion in the dataset included that the schools be public (no private schools or alternative education settings), serve primarily an elementary level

population, and have SET 2.1 data for the 2011-2012 school year. Additionally, it was requested that TIC 3.1 data be provided for those schools that had administered both the SET and the TIC during that year. NCES data from the 2011-2012 school year were merged with the database, providing information regarding the schools' enrollment, Title I status, percentage of students receiving free or reduced lunch, state, percentage of students from ethnic minority backgrounds, state, and locale (e.g., rural, town, suburban, or urban location).

The database provided by ECS contained SET data for 1864 schools spanning a greater range of education levels (preschool through high school), SWPBIS implementation status (pre and post), and location (both within and outside the United States) than was necessary for this study. For information regarding the attributes of the schools in the ECS database and their student populations, NCES-reported data were given preference over SWIS self-reported data given that they were collected specifically for the 2011-2012 school year and contain more detailed information about the demographic composition of the schools. Exceptions to this preference were made for the purposes of filtering data by country and state, as all schools had basic geographic information provided through PBISapps.org/SWIS. The sample was filtered to exclude schools outside of the United States and no schools were duplicated within the dataset.

For the purposes of this study, schools exclusively serving two or more grades between Kindergarten and the 6th grade were classified as being elementary schools, irrespective of whether the school served preschool populations. Schools that included middle and high school grades (7-12+) were excluded, as the "Expectations Taught" and "Management" SET subscales have been found to be difficult to interpret and need

revision above the elementary level (Vincent et al., 2010). In the cases of schools without NCES data, SWIS data were used to determine the grade levels served.

While ECS is able to provide data regarding how many years the schools have been *entering* SWPBIS implementation data using PBISapps.org, no information is available regarding how many years the school has been *implementing* SWPBIS. Information is available regarding the school's SWPBIS implementation status however, indicating whether schools were *pre-* or *post-implementation* at the time of data entry. For this study, it was important that all schools included in the study had data for a SET 2.1 administered "post-implementation," so that SET and TIC scores reflected the level of implementation fidelity in a school actively implementing a SWPBIS program. Therefore, the database was filtered to exclude schools that were pre-implementation or were missing implementation status data, ensuring that all SET 2.1 data were from an assessment administered post-implementation. This excluded schools that had not begun SWPBIS training or were still in the planning stages of SWPBIS development.

After filtering the dataset according to a group of all public elementary schools actively implementing SWPBIS within the United States, there were a total of 1018 schools with SET 2.1 data. Based on SWIS data, it was determined that these 1018 schools were in 361 districts located across 28 states. Of these schools, 947 had NCES data; however, not all schools with NCES data had complete demographic information. Though the NCES data include locale data regarding the geographic location of schools, they provided greater specificity than was needed for this study. Therefore, the urban locale data from NCES were simplified from 12 subtypes to the four main types of locations (e.g., City, Suburb, Town, Rural). The sample was composed of schools

serving a large range of student body sizes (44-1258 students), predominantly included schools eligible for Title I (79.08% of the sample), and included schools across rural, town, suburban, and city locales. Refer to Table 2 for complete demographic characteristics of the sample, including data on schools' percentages of low-income students (i.e., students receiving free and reduced lunches) as well as the schools' racial/ethnic background composition.

Within the ECS dataset, the majority of schools had earned passing scores on the SET. Of the 1018 elementary schools in the sample, 86.54% (881 schools) had met the 80/80 SET criteria, demonstrating that they had achieved an 80% overall score and 80% on the "Behavioral Expectations Taught" subscale. The average overall implementation score on the SET was 91.94% (SD = 8.23%), with a range from 39% to 100%. The middle 50% of schools in this data set scored between 88.00% and 98.00%. These statistics indicate that the vast majority of schools in this data set were successfully implementing SWPBIS according to the SET's established scoring methods.

In addition, within the larger sample of schools with 2011-2012 SET data, a subset of 492 schools also had TIC 3.1 data for that same year. Of these 492 schools, 53.66% of them had met the 80% overall score criteria for SWPBIS implementation fidelity on the TIC, with the mean overall TIC implementation score being 80.16% (SD = 13.26%) of the total possible points. The lowest and highest TIC scores within the dataset were 36.36% and 100%, respectively, with the interquartile range being 72.73% to 90.901%. Refer to Table 3 for a descriptive summary of the dataset's implementation scores on the SET and TIC.

Table 2. Demographic Data for ECS Database

	Elementary Schools with SET 2.1 Data
Number of Schools	1018
Number of Schools with NCES Data	947
Number of Schools with TIC 3.1 Data	492
Student Enrollment (*n* = 943*)	
M (SD)	461.62 (183.93)
Range	44 - 1258
Students Receiving Free and Reduced Lunch (*n* = 943*)	
M (SD)	56.07% (25.85%)
Range	0% – 99.66%
Schools Eligible for Title I (N = 1018**)	
Yes	79.08%
No	13.65%
Missing Data	7.27%
Locale (N = 1018***)	
Rural	22.69%
Town	12.57%
Suburb	26.42%
City	31.34%
Missing Data	6.97%
Student Race/Ethnicity (*n* = 943*)	
American Indian/Alaskan Native *M (SD)*	1.50% (5.99%)
Asian American *M (SD)*	4.56% (8.51%)
Black *M (SD)*	19.50% (25.51%)
Native Hawaiian/Pacific Islander *M (SD)*	0.25% (0.61%)
Hispanic *M (SD)*	18.47% (21.90%)
White *M (SD)*	51.64% (31.47%)
More than One *M (SD)*	4.09% (3.65%)

Note: All schools are at the elementary level.
*Missing data on 75 schools out of the 1018 total.
**Missing data on 74 schools; accounted for in calculation of percentages.
***Missing data on 71 schools; accounted for in calculation of percentages.

Table 3. Dataset Descriptive Statistics For Implementation Scores

	Established SET 2.1 Score	TIC 3.1 Score
	N = 1018	*n* = 492
m (SD)	91.94% (8.23%)	80.16% (13.26%)
Range	39% - 100%	36.36% - 100%
Interquartile Range	88.00% - 98.00%	72.73% - 90.91%

Materials

SET

The SET (Version 2.1; Sugai et al., 2005) as described above was also used in this second study. SET 2.1 data from the Education and Community Supports database was used both at the overall level for the existing two methods of scoring and at the item level to test the four new scoring approaches: unweighted, re-weighted, unweighted with dropped items, and re-weighted with dropped items.

Team Implementation Checklist (TIC)

The Team Implementation Checklist (TIC) is a 22-item rating scale, developed to be completed by the SWPBIS team quarterly throughout the school year (Tobin, 2006). The TIC was developed by Sugai, Horner, and Lewis-Palmer (2002), and is currently in Version 3.1 (Sugai et al., 2011). To complete the TIC, the school's PBIS team discusses and rates each item while a designated member completes a single TIC form that is then submitted to the school's SWPBIS coach or coordinator. The items on the TIC include implementation factors across seven categories of SWPBIS procedures, including: 1) establish commitment; 2) establish and maintain team; 3) self-assessment; 4) establish

school-wide expectations: prevention systems; 5) classroom behavior support systems; 6) establish information systems; and 7) build capacity for function-based support. Each item across these categories is rated by the team as being at one of three levels of implementation, with a point value assigned to each level as follows: "Not Started" worth 0 points, "In Progress" worth 1 point, and "Achieved" worth 2 points. Tobin, Vincent, Horner, Rossetto Dickey, and May (2012) note that in addition to assessing SWPBIS Tier I, the TIC also assesses readiness and preparations for Tier II and Tier III applications (within "Build capacity for function-based support").

In considering the concurrent validity of the SET in comparison to the TIC, the validity and reliability of the TIC must also be evaluated to establish its relevance to the measurement of implementation fidelity as well as its technical adequacy (Messick, 1988). TIC Version 3.1 was developed in August 2012, providing elaborated and more specific standards for each of the 22 items included in TIC 3.0, so data from either version would be comparable; however, previous versions had as many as 26 items and therefore are not useful in considering the current validity of the TIC. A recent unpublished study by Tobin and colleagues (2012) of 893 schools looked at the TIC 3.0's concurrent validity with the BoQ 2.0, finding a moderate Pearson correlation ($r = .59$; $p \leq 0.01$; 2-tailed). Using a sample of 3,408 schools, the TIC 3.0 was found to have high internal consistency ($\alpha = .91$).

The validity of the TIC's scoring has also been studied in relation to the BoQ Revised, another commonly used tool for evaluating SWPBIS implementation fidelity. As discussed above, the TIC is most commonly scored by finding a percentage of total possible points by adding together the total number of items that the SWPBIS team has

rated as either "achieved" (2 points) or "in progress" (1 point). Vincent and Tobin (2012) examined whether including only those items rated "achieved" (rather than both "achieved" and "in progress") would make the TIC a more valid indicator of SWPBIS implementation. After scoring the TIC both ways, the scores were compared to the BoQ, looking at whether schools achieving 80% of the TIC points or 80% of the TIC items rated "achieved" met the implementation criteria of 70% on the BoQ. The study reaffirmed that the TIC has strong concurrent validity with the BoQ, and found that the two different approaches to scoring did not make a statistical difference in the TIC's alignment with the BoQ. Based on these results, for the purposes of this study, the TIC total score was calculated by finding the percentage of total points earned by items scored as either "achieved" or "in progress."

CHAPTER 4

RESULTS

Study 1

The first study aimed to identify which items on the SET are considered to be most critical for sustainable and successful SWPBIS implementation by asking the most prolific experts in the field to rate the importance of each SET item. Based on the SET Validation Survey, the experts' ratings were collected and evaluated to answer the first research question: whether their ratings suggest alternative methods of scoring the SET that might improve its content validity.

Participant Demographics

Of the 58 SWPBIS experts invited to participate in the SET Validation Survey, a total of 20 individuals participated and completed the survey. Participation was anonymous and all experts took the survey in English. The sample included primarily individuals with doctoral degrees (85%), but also individuals with master's degrees (10%) and with both a doctoral degree and law degree (5%). Approximately half (55%) of the experts identified themselves as professors, while 20% identified themselves as researchers at SWPBIS implementation agencies. The remaining 25% work with SWPBIS as researchers, evaluators, or practitioners across a variety of public and private settings. The SWPBIS experts had, on average, 15.43 years of study in the field, with a range of 6 to 34 years. The middle 50% of the experts had between 11.5 and 34 years of experience. With respect to race, the sample was very homogeneous, self-identifying as

95% White and 5% Black. See Table 4 for complete demographic details of survey participants.

Table 4. Demographic Characteristics of Survey Participants

Gender	
Male	50% (*n* = 10)
Female	50% (*n* = 10)
Years of Experience with SWPBIS	
Mean	15.43
Range	6 - 34
Interquartile Range	11.5 to 34
Education Level	
Master's Degree	10% (*n* = 2)
Doctoral Degree	85% (*n* = 17)
Doctoral Degree & Law Degree	5% (*n* = 1)
Occupation	
Professor	55% (*n* = 11)
Researcher at SWPBIS Implementation Agency	20% (*n* = 4)
Other[a]	25% (*n* = 5)
Ethnicity	
Not Hispanic or Latino	100% (*n* = 20)
Race	
African American/Black	5% (*n* = 1)
White	95% (*n* = 19)

Note: N= 20

[a]Includes University Researcher, Evaluator at State-Level SWPBIS Project, Researcher at Independent/Nonprofit Institute, and dual roles as Professor/Researcher at SWPBIS Agency and School-based Practitioner/Educational Consultant.

SET Validation Survey Results

Each participant completed the SET Validation Survey, including rating all 28 SET items as well as giving comments on the elements of SWPBIS they believe are

missing from the existing SET (Appendix F) and their preferred method of assessing SWPBIS implementation fidelity (Appendix G). The complete ratings given by the 20 experts are included in Appendix I.

The range of responses and frequency of each rating (1 through 5) received by each SET item on the SET Validation Survey was considered to determine whether there was sufficient polarization in the experts' ratings to warrant a follow-up survey round. Refer to Table 5 and Table 6 for the response frequency distributions and range of ratings, respectively. The goal of this survey was to obtain mean ratings of the relative importance of each SET item to sustainable SWPBIS implementation. Therefore, it was important that any items for which there was a clear polarized split in ratings be reconciled to obtain a mean rating that represented as close to a consensus as possible. An examination of the SET Validation Survey's frequency distribution did not reveal any significant polarized response patterns with a lack of mid-range ratings. Rather, ratings were found to either cluster closely (e.g., a range of 4 to 5 ratings) or scatter more widely (e.g., a range from 1 to 5), but with each rating being represented. Therefore, no need was found for a second round of expert surveying.

Table 5. Distribution of Ratings for Each SET Item on the SET Validation Survey

	1 Not at All Important for Sustainable Implementation of SWPBIS	2 Helpful, but not Necessary for Sustainable Implementation of SWPBIS	3 Important, but not Necessary, for Sustainable Implementation of SWPBIS	4 Very Important, but not Critical, for Sustainable Implementation of SWPBIS	5 Critical for Sustainable Implementation of SWPBIS
SET Item A1	-	-	5%	15%	80%
SET Item A2	-	-	15%	30%	55%
SET Item B1	-	-	-	5%	95%
SET Item B2	-	10%	5%	20%	65%
SET Item B3	-	10%	10%	20%	60%
SET Item B4	-	-	5%	25%	70%
SET Item B5	-	-	-	15%	85%
SET Item C1	-	-	5%	35%	60%
SET Item C2	-	-	15%	40%	45%
SET Item C3	-	-	15%	25%	60%
SET Item D1	-	-	-	35%	65%
SET Item D2	-	-	10%	20%	70%
SET Item D3	5%	25%	30%	20%	20%
SET Item D4	5%	15%	35%	15%	30%
SET Item E1	-	5%	10%	25%	60%
SET Item E2	-	5%	5%	20%	70%
SET Item E3	-	5%	25%	10%	60%
SET Item E4	-	-	10%	15%	75%
SET Item F1	-	10%	25%	40%	25%
SET Item F2	-	-	15%	40%	45%
SET Item F3	-	5%	40%	25%	30%
SET Item F4	-	10%	25%	45%	20%
SET Item F5	-	-	5%	15%	80%
SET Item F6	-	5%	15%	20%	60%
SET Item F7	-	10%	10%	45%	35%
SET Item F8	-	-	10%	30%	60%
SET Item G1	-	5%	30%	35%	30%
SET Item G2	-	15%	35%	30%	20%

Note: N = 20

Table 6. SET Validation Survey Descriptive Statistics

SET Item	Expert Survey Mean Rating	Median Rating	Mode	Range
SET Item A1	4.75	5	5	3 to 5
SET Item A2	4.40	5	5	3 to 5
SET Item B1	4.95	5	5	4 to 5
SET Item B2	4.40	5	5	2 to 5
SET Item B3	4.30	5	5	2 to 5
SET Item B4	4.65	5	5	3 to 5
SET Item B5	4.85	5	5	4 to 5
SET Item C1	4.55	5	5	3 to 5
SET Item C2	4.30	4	5	3 to 5
SET Item C3	4.45	5	5	3 to 5
SET Item D1	4.65	5	5	4 to 5
SET Item D2	4.60	5	5	3 to 5
SET Item D3[a]	3.25	3	3	1 to 5
SET Item D4[a]	3.50	3	3	1 to 5
SET Item E1	4.40	5	5	2 to 5
SET Item E2	4.55	5	5	2 to 5
SET Item E3	4.25	5	5	2 to 5
SET Item E4	4.65	5	5	3 to 5
SET Item F1[a]	3.80	4	4	2 to 5
SET Item F2	4.30	4	5	3 to 5
SET Item F3[a]	3.80	4	3	2 to 5
SET Item F4[a]	3.75	4	4	2 to 5
SET Item F5	4.75	5	5	3 to 5
SET Item F6	4.35	5	5	2 to 5
SET Item F7	4.05	4	4	2 to 5
SET Item F8	4.50	5	5	3 to 5
SET Item G1[a]	3.90	4	4	2 to 5
SET Item G2[a]	3.55	3.5	3	2 to 5

[a] Indicates items falling in the lowest quartile.

Mean Ratings

For each item of the SET, the importance ratings of each of the twenty participating experts were averaged together to obtain a mean importance rating. Mean ratings, along with the median, mode, and range of ratings for each item are provided in Table 6. While individual experts' ratings of SET items ranged from *1* ("Not at All

Important for Sustainable Implementation of SWPBIS") to *5* ("Critical for Sustainable Implementation of SWPBIS"), the mean ratings ranged from 3.25 to 4.9. The majority of the SET items received mean ratings of *4* ("Very Important, but not Critical for Sustainable Implementation of SWPBIS"), indicating that the experts' agreed that most of the existing SET items are very important or critical for sustainable SWPBIS implementation. These mean ratings were obtained to inform the development of a new weighted scoring method for the existing SET items.

Identifying Lowest-Rated Items

In addition to exploring the mean ratings given to each of the SET items by experts, this study also aimed to consider whether the experts' ratings might be used to refine the SET to include only those items that were considered most critical to sustainable implementation. As discussed above, the experts' mean ratings ranged from 3.25 to 4.95, with the majority of items falling above 4.00, or very important. This cut point, 4.00, also corresponded closely with the lower quartile of the mean ratings, with 25% (or 7 items) falling below 4.01. The items falling in the lower quartile are indicated on Table 6. The lowest rated SET items with mean ratings below 4.01 are listed in Table 7 and include items relating to crisis planning, school improvement goals, team membership, budgeting, and external support. For the purposes of this study, these seven items were determined to be least critical to sustainable SWPBIS development and implementation, and therefore were excluded from two of the novel scoring approaches as described in the following sections.

Table 7. Lowest Quartile SET Items to Exclude

Item	Mean Rating	SET Item Wording
D3	3.25	Is the documented crisis plan for responding to extreme dangerous situations readily available in 6 of 7 locations?
D4	3.50	Do 90% of staff asked agree with administration on the procedure for handling extreme emergencies (stranger in building with a weapon)?
F1	3.80	Does the school improvement plan list improving behavior support systems as one of the top 3 school improvement plan goals?
F3	3.80	Does the administrator report that team membership includes representation of all staff?
F4	3.75	Can 90% of team members asked identify the team leader?
G1	3.90	Does the school budget contain an allocated amount of money for building and maintaining school-wide behavioral support?
G2	3.55	Can the administrator identify an out-of-school liaison in the district or state?

Development of Novel SET Scoring Methods

Unweighted SET

As discussed by Wang and Stanley (1970), the most efficient method of weighting scale items is most frequently to allow natural or unit weighting, in which each item is given equal weight. While the established SET scoring method did not explicitly weight items, each item had an implicit weight determined by how many items were in each of the seven subscales. As subscale means and then an overall mean score of the seven subscales were determined, items were indirectly given weights ranging from $1/56^{th}$ of the overall score to $1/14^{th}$ of the overall score. Therefore, the first novel approach to scoring the SET simply equalized the weight of all items. To that end, an overall unweighted SET score was calculated by adding up the total number of points

earned across all 28 items, and dividing them by the total number of points possible. The total number of points possible was two points for each of the 28 items, for a total of 56 points. The score as a percentage of total points is calculated as follows:

$$\textbf{Unweighted Score} = \frac{\Sigma(\text{Score}_{\text{ItemA1.}} + \text{Score}_{\text{ItemA2}} + \ldots + \text{Score}_{\text{ItemG2}})}{56} * 100$$

Reweighted SET

For the second novel SET scoring approach, the mean expert ratings (Table 6) were used as a multiplier to weight each of the corresponding 28 SET items according to their relative importance to sustainable SWPBIS. Therefore, each SET item was worth the number of points earned (with a maximum of 2 points per item) multiplied by the mean expert rating for that item. The school's total points earned were then divided by the new total possible points, 240.4 - found by finding the sum of all 28 items when the maximum points a school can earn per item (2) was multiplied by the mean expert rating received by each item. The reweighted SET scores were calculated as a percentage of total points using the following formula and mean ratings found in Table 6:

$$\textbf{Reweighted Score} = \frac{\Sigma[(\text{Score}_{A1} * \text{MeanRating}_{\text{ItemA1}}) + (\text{Score}_{A2} * \text{MeanRating}_{\text{ItemA2}}) + \ldots + (\text{Score}_{G2} * \text{MeanRating}_{\text{ItemG2}})]}{240.4} * 100$$

Unweighted SET with Dropped Items

Using the experts' mean ratings, the unweighted SET scoring method described above was modified to exclude the seven items in the lower quartile (Table 7) with respect to relative importance to SWPBIS implementation. Those seven items were dropped from the sum of total points earned and from the total possible points. The unweighted SET equation was modified to exclude the dropped items and adjusted with a denominator of 42, the new total possible points.

Reweighted SET with Dropped Items

Finally, the reweighted SET equation was modified to exclude the seven dropped items (Table 7). The seven items were excluded from the calculation of the numerator and the total possible points sum was adjusted accordingly. The maximum possible points after the bottom quartile items were dropped was calculated to be 189.3.

Based upon these four methods, SET item data from the ECS database were used to recalculate four new SET scores for all 1018 schools. The old and new methods of scoring were summarized using descriptive statistics, shown in Table 8. Based on their mean scores, there was little variation with regard to clinical differences in the total scores yielded. In other words, while the mean score for the established SET was 91.94%, the novel scoring methods all fell within a single percentage point, which would likely have little impact on a school's interpretation of its results. Analyses were conducted in Study 2 to look more closely at whether these changes in scoring were related to improvements in the SET's concurrent validity with the TIC.

Table 8. Descriptive Statistics for Established and Novel SET Scoring Methods

	M (SD)	Range
Established SET	91.94% (8.23%)	39% - 100%
Unweighted SET	91.85% (8.12%)	34% - 100%
Reweighted SET	91.98% (8.13%)	34% - 100%
Unweighted SET with Dropped Items	92.76% (8.49%)	36% - 100%
Reweighted SET with Dropped Items	92.73% (8.52%)	36% - 100

Note: N = 1018

Factors Missing from the SET

In addition to providing quantitative ratings of each SET item, the experts were asked to identify any SWPBIS implementation factors or components they judged to be missing from the SET. Summarizing the components into categories (Appendix F), the most frequent suggestion (noted by four experts) was that the SET was lacking items to assess classroom-level PBIS systems and plans. The second most frequently noted missing element (noted by three experts) was a consideration for the cultural relevance and fit of the SWPBIS program and whether the program was responsive to the cultures and backgrounds of the school community. Two experts each noted that the SET might benefit from breaking down elements into more specific goals and is missing items related to family engagement/involvement in SWPBIS as well as secondary and tertiary supports. Other SWPBIS components identified as missing from the SET included: bullying prevention, direct observation of staff SWPBIS practices, quality indicators, supervision, faculty buy-in, and integration of behavioral education throughout the year.

Study 2

Having explored the content of the SET and developed four new methods of scoring the SET based on the SET Validation Survey in Study 1, the goal of Study 2 was to investigate whether changing the weighting of the SET items or excluding the relatively less important items affected the concurrent validity of the SET with the TIC. For these analyses, this study used both the 80/80 scoring standard developed by Horner and colleagues (2004) and the 80% total SET score standard used in research (Cohen et al., 2007; Horner et al., 2009), as well as the four novel scoring approaches established in

Study 1. It is important to note that the validated SET scoring methods include one dichotomous measure (pass/fail) and one continuous measure (overall score). The four new methods of scoring the SET were scored similarly to the overall implementation scoring method of scoring the TIC, finding a percentage of overall implementation by dividing the total number of points earned by the total possible points. Therefore, the new methods are each scored on a continuous scale, and were held to the same standards as the SET and the TIC – 80% or higher to indicate successful SWPBIS implementation.

Similar to the SET scoring, the TIC may also be scored in two ways. The established method of scoring the TIC (see Sugai et al., 2005 and Vincent & Tobin, 2012) is to obtain a total score based on the number of points earned out of the total possible points, such that the TIC score may be used as a continuous variable. In practice within schools, however, implementation fidelity tools are frequently used in a pass/fail manner. The TIC overall score can be transformed into a dichotomous variable indicating whether the school passed (meets 80% or higher criteria) or failed (earned fewer than 80% of the total possible points). This study used both methods of scoring the TIC, allowing an exploration of the utility of each method compared to the SET. While the dichotomous scoring approach lends itself to ease of understanding and direct comparison to the SET dichotomous outcome metric, using the TIC as a score on a continuous scale allows for greater sensitivity in analyses. Therefore, the concurrent validity of the SET and TIC scores was evaluated based on six methods of scoring the SET and two methods of scoring the TIC.

Concurrent Validity of Overall Scoring Methods

Bivariate Pearson correlations were calculated to determine relationships between the continuous variables, including the established SET overall score, the four novel SET scoring methods, and the overall TIC implementation score. As might be expected given that all of the SET and TIC metrics included in the analysis are methods of assessing SWPBIS implementation, positive and significant correlations were found between the established SET overall scoring method and each of the novel SET scoring approaches, as well as the TIC overall score (see Table 9). Among the five SET scoring methods included in this analysis, the correlations were strong (ranging from $r(1016) = .881$ to $r(1016) = .998$, $p < .001$. This indicates that as the SET scoring varied with respect to item weighting and exclusions, there remained a strong relationship between the scores.

An analysis of the pattern of correlation scores between SET versions in Table 9, however, showed a trend such that the unweighted and reweighted SET scoring methods correlated most strongly with the established SET method (significant correlations above .960) while the unweighted and reweighted methods with dropped items were slightly less correlated with the established SET ($.881 \leq r \geq .882$), though still indicating a strong relationship. Worth noting, the scoring approaches using unweighted and reweighted SET items and dropping those items rated least important were perfectly correlated with each other ($r(1016) = 1.000$, $p < .001$). These trends suggest that dropping items, not type of weighting, was associated with change in the SET scores.

Table 9. Pearson Correlations Between Continuous SET and TIC Scoring Methods

	Established SET	Unweighted SET	Reweighted SET	Unweighted SET with Dropped Items	Reweighted SET with Dropped Items	TIC Overall
Established SET	----					
Unweighted SET	.965***	----				
Reweighted SET	.960***	.998***	----			
Unweighted SET with Dropped Items	.882***	.936***	.954***	----		
Reweighted SET with Dropped Items	.881***	.935***	.954***	1.000***	----	
TIC Overall	.360***	.377***	.382***	.401***	.400***	----

Note: For SET scoring methods, N = 1018. For correlations calculated with TIC overall score, N = 492.
***Correlation is significant at the $p < 0.001$ level (2-tailed).

Considering the relationships of the established SET scoring method and the four novel SET scoring methods with the TIC overall score, each method of scoring the SET was found to have a significant moderate and positive relationship to the TIC. The established SET and TIC overall score had the lowest correlation compared to the new SET scoring approaches ($r(490) = .360$, $p < .001$). This moderate correlation makes sense within the context of the larger sample in which the mean established SET score

(N=1018) was 91.94%, while the mean TIC overall score (N=492) was 80.16% (see Table 3). Within the subsample of 492 schools which had both SET and TIC data, the overall SET score and TIC score were compared using a paired samples *t*-test. SET scores were found to be significantly higher than the TIC ($t(491)= 19.138$, $p < .001$). Therefore, this difference between SET and TIC, such that schools tend to score higher on the SET than the TIC, is consistent throughout the subsample of schools with SET and TIC data.

The four new methods of scoring the SET also demonstrated positive moderate correlations with the TICs, as seen in Table 9. The novel scoring approaches appeared to cluster such that the unweighted SET ($r(490) = .377$, $p < .001$) and reweighted SET ($r(490) = .382$, $p < .001$) were somewhat more strongly related to the TIC overall score than the established SET score ($r(490) = .360$, $p < .001$). The unweighted SET with dropped items ($r(490) = .401$, $p < .001$) and reweighted SET with dropped items ($r(490) = .400$, $p < .001$), correlated with the TIC even more strongly than the scoring methods that included all 28 items.

Based on the observed trends that dropping items seemed to have the greatest impact on the SET's concurrent validity both with the established SET and the TIC, the correlations between the SET and the TIC were compared. To test whether the new methods of scoring the SET statistically improved the concurrent validity of the SET over the established SET scoring method, an "asymptotic *z*-test" was used (Lee & Preacher, 2013), which allows for comparison of correlations that share a variable between them (in this case, the TIC overall score). This test compared the correlations between each novel SET scoring approach/TIC overall score and the established SET Score/TIC overall score

to determine whether the novel scoring methods had statistically higher concurrent validity with the TIC overall score. Lee and Preacher's web software first converts each correlation into a *z* score using Fischer's *z* transformation and then calculates the difference between them by comparing them against the correlation of the novel SET scoring approach with the established SET Score. The results, as shown in Table 10, indicate that while the unweighted and reweighted SET scoring approaches do not differ significantly from the established SET overall scoring approach, the novel scoring methods in which low-rated items were dropped were significantly different. Within this dataset, the overall scores produced by the unweighted SET with dropped items method (z = -2.03, $p < .05$) and the reweighted SET with dropped items method (z = -1.98, $p < .05$) were both significantly different from the established SET scoring approach. Therefore, the two scoring methods with dropped items were both found to be significantly stronger predictors of TIC scores than the established SET or either the unweighted or reweighted SET scores.

Table 10. Differences in Concurrent Validity with the TIC Overall Score between the Established and Novel SET Scoring Methods

	Unweighted SET	Reweighted SET	Unweighted SET with Dropped Items	Reweighted SET with Dropped Items
Established SET	z = -1.53	z = -1.86	z = -2.03*	z = -1.98*

*2-tailed p < 0.05

Concurrent Validity of Dichotomous Scoring Methods

In addition to being scored as with a total overall implementation score (established SET scoring method), the SET was developed to also have a pass/fail designation based on an 80% goal for overall implementation as well as 80% implementation within the "Behavioral Expectations Taught" category (Horner et al., 2004). This pass/fail designation is a dichotomous measure, entered into SPSS with ordinal properties such that "0" represents a failure to successfully implement SWPBIS and "1" represents success. Similarly, the TIC score can be conceptualized as a dichotomous measure (based on an 80% overall implementation score), particularly in practice when it might be discussed as pass/fail rather than an exact score. Therefore, the next stage of analyses focused on evaluating the relationships between the continuous TIC and SET scoring methods with the dichotomous SET and TIC scoring data.

First, 2-tailed bivariate Spearman correlations were conducted to examine the relationship between the SET 80/80 Criterion and both TIC scoring methods: overall score and dichotomous. Spearman correlations were used to find associations between the dichotomous variables without interval or ratio properties (Choi et al., 2010). As shown in Table 11, the SET 80/80 Criterion, though significantly correlated with the TIC 80% criterion ($r(490) = .165$, $p < .001$) and the TIC overall score ($r(490) = .223$, $p < .001$), both are weak correlations compared to the moderate Pearson r's between the Established and novel SET scoring approaches and the TIC overall score. The TIC overall score and the TIC 80% criterion were strongly and positively correlated ($r(490) =$

.875, $p < .001$), which makes sense given that the dichotomous TIC outcome is a simplified snapshot of the TIC overall score.

Table 11. Spearman Correlations Between Dichotomous SET Score and TIC

	SET 80/80 Criterion	TIC 80% Criterion	TIC Overall
SET 80/80 Criterion	----		
TIC 80% Criterion	.165***	----	
TIC Overall	.223***	.865***	----

Note: $n = 492$
**Correlation is significant at the $p < .001$ level (2-tailed).

Next, two additional Spearman correlations were run, comparing each of the continuous SET scoring methods (the established SET and the four novel scoring approaches) to the dichotomous TIC 80% criterion and the dichotomous SET 80/80 Criterion. Each of the continuous SET scoring approaches correlated positively and significantly, but weakly with the TIC 80% criterion, as shown in Table 12, with Spearman's *r*'s below .300. These weak correlations are indicative of the discrepancy between the statistically higher mean scores across SET versions compared to the TIC, as well as the loss of sensitivity that occurs when continuous scale data are transformed into a dichotomous variable. Though the SET 80/80 Criterion also loses sensitivity in the same way, moderate positive and significant Spearman correlations were observed between it and the continuous SET scoring methods (see Table 13).

Table 12. Spearman Correlations Comparing the TIC 80% Criterion to the Established and Novel SET Scoring Methods

	Established SET	Unweighted SET	Reweighted SET	Unweighted SET with Dropped Items	Reweighted SET with Dropped Items
TIC 80% Criterion	.262***	.288***	.285***	.292***	.287***

Note: $n = 492$
*** Correlation is significant at the $p < .001$ level (2-tailed).

Table 13. Spearman's Correlations Comparing SET 80/80 Criterion to the Established and Novel SET Scoring Methods

	Established SET	Unweighted SET	Reweighted SET	Unweighted SET with Dropped Items	Reweighted SET with Dropped Items
SET 80/80 Criterion	.548***	.557***	.558***	.553***	.554***

Note: N = 1018
*** Correlation is significant at the $p < .001$ level (2-tailed).

Overall, the weakest correlations were found between the SET 80/80 Criterion and the TIC 80% criterion variables. Additionally, the dichotomous SET and TIC measures were found to have weak concurrent validity in their respective relationships to the established SET and novel SET scoring methods, suggesting that the dichotomous measures are not a useful metric for accurately predicting overall scores. Compared to the stronger moderate correlations found between the continuous SET scoring methods and the overall TIC score, the dichotomous measures have less sensitivity. In simplifying

a full scale implementation score into a pass/fail measure, too much information is lost. Therefore, considering all of the analyses conducted in Study 2, these results suggest that the continuous SET scoring methods have greater concurrent validity with the TIC overall score than the dichotomous SET 80/80 criterion. More specifically, the unweighted SET with dropped items and the reweighted SET with dropped items demonstrated the strongest concurrent validity with the TIC overall score than the existing or other novel SET scoring methods.

CHAPTER 5

DISCUSSION

The goal of this study was to examine the content and concurrent validity of the SET for evaluating SWPBIS implementation fidelity. First, the content and structure of the SET was evaluated by asking prolific experts in the field to rank the SET items with regard to their relative importance to sustainable SWPBIS implementation. Based on the experts' mean ratings on the SET Validation Survey as well as theory related to item weighting, four novel methods of scoring the SET were developed: unweighted SET, reweighted SET, unweighted SET with dropped items, and reweighted SET with dropped items. The novel and existing SET scoring methods were then correlated with the TIC in Study 2, to examine the concurrent validity of the SET with a second measure of SWPBIS implementation. In doing so, the current study aimed to consider whether revisions to the scoring and structure of the SET might improve its validity and utility for both practitioners and researchers.

Based upon the SET Validation Survey results and subsequent analyses of the SET and the four novel SET scoring approaches, this study suggests that the SET's concurrent validity with the TIC is highest when the lowest expert-rated items are excluded from the overall score calculation. The four novel SET scoring approaches developed varied along two dimensions: how items were weighted and whether the lowest rated items were dropped. Each approach was found to have positive and significant moderate correlations with the TIC. While changing the weights of items to both unweighted (equal weighting) and reweighted (according to expert mean ratings) SETs still produced significant correlations with the TIC, their concurrent validities with

the TIC did not differ significantly from that of the established SET score. In other words, based on the asymptotic *z*-test, there was no significant change in the SET's concurrent validity with the TIC over the established SET scoring method either when all 28 items were given equal weight or when all 28 items were reweighted with their mean expert rating. The only significant improvements in SET scoring came from dropping the seven items whose mean expert ratings fell within the lowest quartile. By dropping these items from both the unweighted and reweighted SET versions, concurrent validity with the TIC was significantly improved compared to the original SET scoring method.

Interestingly, while this study found that the SET had greatest concurrent validity with the TIC when it was scored without the seven lowest-rated items, statistically, there was no difference between whether the items had been weighted equally or reweighted. Given the near equivalence of the Pearson *r*'s yielded by these two methods and applying Occam's Razor's emphasis on parsimony in explanations (Feuer, 1957), it is recommended that using the natural or unweighted approach to scoring — combined with the dropped items — is the most appropriate and efficient method of scoring the SET. Within the context of applied science, Occam's Razor and the law of parsimony suggests that when considering multiple hypotheses or theories that provide similar results and are similarly supported, more complex hypotheses based on a greater number of postulates or assumptions should be discarded in favor of more parsimonious hypotheses that rely on fewer underlying assumptions (Ferguson, 1954; Feuer, 1957). The greater number of assumptions underlying a hypothesis, the more likely it is that one or more of those postulates may be disproved or flawed. Therefore, given that the unweighted SET with dropped items and reweighted SET with dropped items had similar concurrent validity

with the TIC and demonstrated a significant (though small) improvement over the established SET scoring, the unweighted method is the more parsimonious – relying on fewer assumptions about the relative importance or weight of each item on the SET.

This study's suggestion that the unweighted SET with dropped items is the more parsimonious method of scoring the SET is consistent with the literature on item weighting, suggesting that the process of obtaining item weights and applying them to a scoring system rarely results in significantly improved results compared to equal unit weighting. As explained by Wang and Stanley (1970), "Although differential weighting theoretically promises to provide substantial gains in predictive or construct validity, in practice these gains are often so slight that they do not seem to justify the labor involved in deriving the weights and scoring with them" (p. 664). This study, therefore, serves both to validate Wang and Stanley's conclusion that natural weighting is best with respect to the SET and to suggest that the most parsimonious method of scoring the SET is simply find a total percentage of implementation points earned.

Though the experts' mean ratings were not useful in developing a novel weighting system for scoring the SET, they were valuable for determining which of the 28 SET items were least critical for sustainable implementation. Based on the SET Validation Survey results, seven items fell within the lowest quartile, with mean ratings less than 4.01, corresponding to average ratings below the rating "Very Important, but not Critical, for Sustainable Implementation of SWPBIS". The seven items (as shown in Table 7), included topics related to having a crisis plan displayed in the school (Item D3), staff knowledge of the crisis plan (D4), documented prioritization of a schoolwide behavior improvement goal (F1), representative team membership (F3), team consensus on team

leader (F4), allocated budget for SWPBIS (G1), and out-of school liaison support (G2). The items remaining are consistent with empirical literature on critical elements of SWPBIS. For example, Vincent and Tobin (2011) summarized that the most necessary factors contributing to successful SWPBIS implementation are: deliberate teaching of behavioral expectations; establishment and use of reinforcement procedures for expected behaviors; consistent consequences for behavioral violations; active "monitoring of student behavior in all school settings;" and collecting and using data on student behavior to problem solve and make decisions. Though the items excluded from the new SET scoring approach may be helpful in developing a SWPBIS system, having a crisis plan or out of school liaison are not core elements of SWPBIS. Although this study indicates that having a crisis plan is not critical for SWPBIS implementation fidelity, this should not be interpreted as meaning that schools do not need crisis plans. Certainly, crisis plans are critically important for school safety and emergency management; however, this study suggests that its presence is not a useful indicator of SWPBIS implementation. Therefore, the unweighted SET with dropped items, provides a more cohesive and prioritized set of SWPBIS elements than the established SET scoring method.

Additionally, this study considered the utility and concurrent validity of the SET's dichotomous pass/fail scoring method, according to which a school must earn an 80% overall score and 80% within the "Behavioral Expectations Taught" subscale to pass (Horner et al., 2004). As might be expected given that the dichotomous measure is a simplified summary of the continuous scores, only weak (though significant) correlations were found between the SET 80/80 and both the Overall TIC and TIC 80% criterion. The four novel SET scoring approaches, as well as the established SET scoring method,

however, all correlated more strongly with both TIC scoring methods. Therefore, this study suggests that in research or in practice when the SET score is being used for comparative analyses or tracking progress over time, it is preferable to use the continuous measures of SWPBIS implementation fidelity - rather than dichotomous measures, including the SET 80/80 Criterion and TIC 80% criterion. Though schools may appreciate the simplicity of the pass/fail standard, the continuous variable allows schools to compare their scores across years and providing more useful information related to level of implementation. When the continuous SET scores are reduced to a pass/fail metric, information is lost and the resulting dichotomous measure of SWPBIS implementation lacks sensitivity and concurrent validity with the TIC.

Implications for Practice

In summary, this research found that dropping the less-critical items and using continuous variable scoring (over both the dichotomous and continuous original SET scoring methods) make a significant difference in the concurrent validity of the SET with the TIC. Based on these findings, practitioners and researchers should be able to easily integrate the novel scoring method into their future use of the SET and apply it to previously-collected SET data. Though this study suggests dropping seven items from the SET and removing all weighting from existing items, the unweighted SET with dropped items scoring method does not add any items to the SET. Therefore, schools and research agencies that have already collected SET data will be able to apply the new suggested scoring to their existing data, retroactively.

Though within the ECS dataset there averaged only one point difference between the established SET (M = 91.94%) and the unweighted SET with dropped items (M =

92.76%), on a school-by-school basis, adopting the new scoring method will help schools to be sure that they are assessing and being scored based on the most critical items for sustainable SWPBIS implementation. Furthermore, changing how an overall score is obtained for the SET does not diminish the value of individual item scores in gauging element-specific progress over time or formulating action plans. Finally, using the unweighted SET with dropped items scoring method may save schools and evaluators time in the future, removing the need to look for crisis plans and cutting down time spent interviewing staff members regarding crisis plans and team leadership. This addresses one of the critiques by Cohen et al. (2007) – that the SET interviewing procedures is time consuming and disruptive to student and staff routines. It is important to note that this study does not suggest that schools should not have and monitor crisis plans, but rather suggests that assessment of crisis plans may not be critical to the measurement of SWPBIS implementation fidelity.

Limitations

Within this research project, there were several limitations. First, one possible limitation was self-selection bias among those experts who chose to participate in the SET Validation Survey. This, however, did not appear to impact the survey results given that the invited experts were recruited based on their investment in SWPBIS. Another limitation can be found in some of the complications in using an existing database for the analyses in Study 2. While having a dataset with 1018 schools is invaluable, it is important to recognize that those schools that use PBISapps.org to track their SWPBIS data may be skewed to have higher SET and TIC scores. Given that the schools are actively using a web-application to record and track their implementation progress, the

schools are likely very committed to, knowledgeable about, and financially invested in SWPBIS. Furthermore, while schools using PBISapps.org must have the cooperation of a PBIS coach, it is not possible to have direct knowledge of who collected and entered data, or whether the assessments were checked for accuracy when collected or entered. Therefore, while the collection of data is theoretically overseen by individuals who have been trained according to the *SWPBS Implementation Blueprint* (Sugai et al., 2010) and the *Evaluation Blueprint for School-Wide Positive Behavior Support* (Algozzine et al., 2010), it is not possible to obtain specific information regarding the training or experience of the specific school-based personnel charged with conducting the evaluations.

Another possible limitation of Study 2 was that the data were sorted to focus only on schools that were in the "post-implementation" rather than planning and development phase of SWPBIS. By focusing on post-implementation, the SET scores were clustered around higher levels of implementation, with a mean of 91.94% (SD = 8.23%) and an interquartile range of 88% to 98% - well above the field's 80% implementation standard. Including a broader set of pre- and post-implementation schools, with a wider range of implementation scores might be helpful to consider in the future. However, as discussed by Norman (2010), the fact that the SET scores were clustered around high levels of implementation and did not adhere to a normal distribution would not have skewed the Pearson or Spearman correlations calculated in Study 2. Therefore, future analyses might include the pre-implementation data to evaluate how the novel scoring methods might affect scoring at lower levels of implementation.

Within this study and the larger SWPBIS evaluation literature, one of the inherent difficulties is conducting descriptive and statistical analyses based on ordinal data. The items on both the SET and the TIC are ordinal in nature, with points being assigned according to levels of implementation (i.e., "0" for not at all, "1" for partially, and "2" for sufficiently). There exists a lack of consensus among statisticians and researchers in the social sciences, with disagreements over whether ordinal data can be used in statistical analyses; however, ordinal data from rating scales and assessment tools are frequently collected and necessarily analyzed within the social and applied science fields (Choi et al., 2010; Norman, 2010). For the purposes of this study, the individual item scores on the SET 2.1 were entered into SPSS analyses as scale (or interval) variables. Though the items are scored between 0 and 2 points using an ordinal scale and lacking an absolute zero, the treatment of the item scores as scale is consistent with how they have been used in research and practice to obtain an overall implementation score or a pass/fail of the 80/80 standard, respectively (Horner et al., 2004). The TIC item scores were given similar treatment given their similar scoring and applications. Additionally, the Likert-type items on the SET Validation Survey in this study were ordinal in nature, but averaged for the purpose of obtaining mean ratings. Though ordinal data may have limitations as far as the analyses that can be computed with them and no clear consensus on their treatment, this research utilized ordinal data in a manner consistent with the SWPBIS field. Results should be interpreted carefully, but are comparable to results from other studies published on the SET or TIC. Finally, it is worth noting that the overall scores for the SET and the TIC are percentages, and therefore are on a continuous scale with interval properties, appropriate for correlations and descriptive analyses.

Future Directions

While the results of this study suggest that using the unweighted SET with dropped items scoring method improves the concurrent validity of the SET as a measure of SWPBIS implementation fidelity, considering the weighting and the exclusion of items are only two specific means of studying the validity of SWPBIS evaluation. Based on this study, future research should focus on the investigation of the concurrent and predictive validity of the unweighted SET with dropped items compared to other measures beyond the TIC, as well as its reliability, sensitivity to change, possible impacts on schools' decision-making around implementation, and applications at the secondary level or alternative settings. Additional studies should consider whether the 80% implementation standard of the SET should remain the standard, as the SET and novel SET scores had a mean approximately ten percentage points higher than the TIC.

Specifically, future studies should examine the relationship between the unweighted SET with dropped items and the relevant universal components of the SWPBIS Tiered Fidelity Inventory (Algozzine et al., 2014), a new SWPBIS implementation fidelity assessment. This new tool aims to provide a single evaluation tool to measure critical components of SWPBIS at the primary, secondary, and tertiary levels. Validating the novel SET scoring approach with this and other new tools will be important in helping schools evaluate their programs over time and across assessments.

Continued research to improve SWPBIS implementation fidelity measurement should also focus on adding items to assess elements of SWPBIS that are lacking from the SET. For example, the experts surveyed for this study reported that the SET lacked coverage in many areas, including (but not limited to) classroom supports, culturally

responsive implementation, family engagement, tiered supports, and buy-in. These areas are consistent with areas being studied and integrated with newer assessments. For example, Marchant, Heath, and Miramontes (2012) suggest that social validity surveys should be paired with implementation fidelity data collection to be obtained most effectively and efficiently. Future research might consider whether an item related to staff input of SWPBIS' social validity should be included on the SET or another implementation assessment.

SWPBIS and Diverse Student Populations

Perhaps one of the most important directions for future research in the measurement of SWPBIS implementation fidelity is greater consideration of school context and student population diversity in the development and evaluation of SWPBIS programs. Among experts surveyed for this study, cultural relevance of SWPBIS was the second most frequent response for missing components of the SET. Citing U.S. Department of Education statistics about the increasing diversity of students within the public education system, Fallon, O'Keefe, and Sugai (2012) suggest that SWPBIS systems must be developed with consideration for the cultural and behavioral norms of all students, not just the school's dominant cultural groups. They suggest that culturally-responsive SWPBIS systems should involve families and communities in the development and teaching of behavior expectations and the use the students' language or colloquialisms. Such consideration of diversity in SWPBIS includes respect for differences in cultural expectations between administration/faculty and students, minimization of disproportionality in regard to use of office discipline referrals, and ensuring that the school environment and expectations remain "culturally and

contextually relevant" to the students and families (Fallon et al., 2012). Cartledge and colleagues (2001) suggest that culturally responsive discipline must incorporate a deliberately developed ethic for the caring of all students. Currently missing from the SET, TIC, BoQ, and other common tools for measuring SWPBIS implementation fidelity, however, are items explicitly addressing issues of contextual fit with regard to cultural and linguistic diversity. Given that student, faculty, and community buy-in is a critical factor in the sustainability of any school-wide initiative (Marchant et al., 2012), future revisions and development of implementation fidelity measures should consider including documentation of consideration of school diversity in program development.

Additionally, research needs to look more specifically at the impacts and outcomes of SWPBIS with regard to academic and behavioral outcomes for children from culturally and linguistically diverse backgrounds. It has been well-documented that - compared to their white peers - students from cultural and linguistic minority groups have higher rates of office disciplinary referrals (Cartledge et al., 2001) and higher rates of exclusionary punishments (Cartledge et al., 2001; NEA, 2007; Sprague et al., 2012). Black students, specifically, have been found to be more likely to receive more severe and exclusionary punishments than their White peers for equivalent rule-violating behaviors (Pauken & Daniel, 1999; U.S. Department of Education OSEP, 2011). These disproportionate disciplinary practices are concerning not just in regard to equality and fairness for all students, but also since higher rates of office discipline referrals and exclusionary punishments have been correlated with lower rates of high school graduation and juvenile delinquency (Skiba et al., 2002; Tobin & Sugai, 1999). Despite these legislative efforts and programmatic changes targeted toward decreasing

disproportionality in special education and school discipline, the nation continues to have significant levels of disproportionality and inequity in its education system (Albrecht, Skiba, Losen, Chung, & Middelberg, 2012; Council for Children with Behavioral Disorders, 2013). Though SWPBIS has been shown to be effective (e.g., Bradshaw et al., 2010; Chitiyo et al., 2012; Solomon et al., 2012), Sprague and colleagues (2012) point out that "evidence-based" does not mean that a given practice, framework, or program is appropriate or culturally relevant to students across all cultures and backgrounds. Therefore, future research should consider how to measure cultural-relevance and appropriateness in SWPBIS implementation.

Integration with Student Outcomes

It would also be beneficial to consider the integration of SWPBIS behavioral and academic outcomes into implementation measurement. As discussed by Sheridan and colleagues (2009), the development of any measure of fidelity should consider the variance each item might account for in the predicted success of that intervention. Horner et al. (2004) identify the relationship between SWPBIS procedures and student academic and behavioral functioning as one of the fundamental issues in SWPBIS assessment. Therefore, future research might look at how changes in student behavior as observed both directly (systematic observation of students) or indirectly (ODRs) are accounted for by each item or category on the SET.

Vincent and Tobin (2012) suggested that since the goal of SWPBIS implementation is to improve student discipline, it would be useful to incorporate a metric of student disciplinary outcomes into fidelity assessment. Based on research demonstrating that successful SWPBIS implementation has been found to decrease office

discipline referrals (e.g., Cohen et al., 2007; Horner et al., 2009; Scott & Barrett, 2004), the average school implementing SWPBIS with fidelity should demonstrate positive changes in their number of office discipline referrals over the first years of implementation. Incorporating a metric related to office discipline referrals into an implementation fidelity measure, however, presents a challenge with respect to how to quantify school-specific and implementation stage-specific changes in referral rates. For example, while a school might experience significant reductions in office discipline referrals over the first several years of implementation, using change across years might become less meaningful once low rates of referrals have been established. Conversely, schools beginning to implement a system of office discipline data collection for major and minor behavioral offenses may see an increase in ODRs as staff are taught to complete and submit referral forms. Similarly, researchers could attempt to develop an item to be scored based on the percentage of the student body receiving office discipline referrals (i.e., percentage of students *not* responding to the SWPBIS system); however, this might unfairly penalize schools still in the initial stages of implementation or alternative school settings serving populations with greater behavioral or emotional needs.

In conclusion, the results of these studies demonstrate the importance and merit of re-evaluating the validity of assessment tools over time. As the field of SWPBIS has matured and its adoption has become more widespread, so has the expertise and empirical evidence regarding which factors are most critical for successful implementation. At the same time, with the ever-increasing diversity and complexity of student needs, characteristics, and challenges, researchers and practitioners must work together to

identify new areas to consider within SWPBIS development, implementation, and assessment. In proposing a new scoring method for the SET, therefore, this study should be considered one step forward within the context of continual progress and improvement in the area of SWPBIS, laying the groundwork for further research and validation.

REFERENCES

Albrecht, S. F., Skiba, R. J., Losen, D. J., Chung, C. G., & Middelberg, L. (2012). Federal policy on disproportionality in special education: Is it moving us forward? *Journal of Disability Policy Studies, 23,* 14-25.

Algozzine, B., Barrett, S., Eber, L., George, H., Horner, R., Lewis, T., … Sugai, G (2014). *School-wide PBIS Tiered Fidelity Inventory.* OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. Available from: www.pbis.org

Algozzine, B., Horner, R. H., Sugai, G., Barrett, S., Rossetto Dickey, C., Eber, L., …Tobin, T. (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.

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

Arthur, W., Jr., Doverspike, D., & Barrett, G. V. (1996). Development of a job analysis-based procedure for weighting and combining content-related tests into a single test battery score. *Personnel Psychology, 49,* 971-985.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91-97.

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

Bond, G. R., Becker, D. R., & Drake, R. E. (2011). Measurement of fidelity and implementation of evidence-based practices: Case example of the IPS Fidelity Scale. *Clinical Psychology: Science and Practice, 18,* 126-141.

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

Brown-Chidsey, R., & Steege, M. W. (2005). *Response to intervention: Principles and strategies for effective practice.* New York: The Guilford Press.

Burns, M. K., Deno, S. L., & Jimerson, S. R. (2007). Toward a unified response-to-intervention model. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention: The Science and Practice of Assessment and Intervention.* New York: Springer.

Carr, E. G., Dunlap, G., Horner, R. H., Koegel, R. L., Turnbull, A. P., Sailor, W., …Fox, L. (2002). Positive behavior support: Evolution of an applied science. *Journal of Positive Behavior Interventions, 4,* 4-16, 20.

Cartledge, G., & Milburn, J. F. (1995). *Teaching social skills to children and youth: Innovative approaches* (3rd Ed.). Boston: Allyn & Bacon.

Cartledge, G., Tillman, L. C., & Johnson, C. T. (2001). Professional ethics within the context of student discipline and diversity. *Teacher Education and Special*

Education: The Journal of the Teacher Education Division of the Council for Exceptional Children, 24, 25-37.

Chitiyo, M., May, M. E., & Chitiyo, G. (2012). An assessment of the evidence-base for school-wide positive behavior support. *Education and the Treatment of Children, 35,* 1-24.

Choi, J., Peters, M., & Mueller, R. O. (2010). Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation. *Asia Pacific Education Review, 11,* 459-466

Clayton, M. J. (1997). Delphi: A technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology, 17,* 373-387.

Cohen, R., Kincaid, D., & Childs, K. E. (2007). Measuring school-wide positive behavior support implementation. *Journal of Positive Behavior Interventions, 9,* 203-213.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Council for Children with Behavioral Disorders. (2013). CCBD's position summary on federal policy on disproportionality in special education. *Behavioral Disorders, 38,* 108-120.

Daly, E. J., III, Lentz, F. E., & Boyer, J. (1996). The instructional hierarchy: a conceptual model for understanding the effective components of reading interventions. *School Psychology Quarterly, 11,* 369-386.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method.* Hoboken, NJ: John Wiley & Sons, Inc.

EBSCO Publishing [Publication Database]. (2014). Retrieved from: www.ebscohost.com.

Educational & Community Supports (Created by Hoselton, R.). (2014). An investigation of the content and concurrent validity of the School-wide Evaluation Tool [Unpublished raw data]. Provided by: Educational & Community Supports, University of Oregon.

Erchul, W. P., & Martens, B. K. (2002). *School consultation: Conceptual and empirical bases of practice* (2nd ed.). New York: Kluwer Academic/Plenum Publishers.

Fallon, L. M., O'Keeffe, B. V., & Sugai, G. (2012). Consideration of culture and context in school-wide positive behavior support: A review of current literature. *Journal of Positive Behavior Interventions, 14,* 209-219.

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in Counseling and Development, 37,* 184-190.

Ferguson, G. A . (1954). The concept of parsimony in factor analysis. *Psychometrika, 19,* 281-290.

Feuer, L. S. (1957). The principle of simplicity. *Philosophy of Science, 24,* 109-122.

Filter, K. J., McKenna, M. K., Benedict, E. A., Horner, R. H., Todd, A. W., & Watson, J. (2007). Check in/Check out: A post-hoc evaluation of an efficient, secondary-level targeted intervention for reducing problem behaviors in schools. *Education and Treatment of Children, 30,* 69-84.

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice, 19,* 531-540.

Handler, M. W., Rey, J., Connell, J., Their, K., Feinberg, A., & Putnam, R. (2007). Practical considerations in creating school-wide positive behavior support in public schools. *Psychology in the Schools, 44,* 29-39.

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

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

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

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

Horner, R. H., Todd, A., Lewis-Palmer, T., Irvin, L., Sugai, G., & Boland, J. (2004). The school-wide evaluation tool (SET): A research instrument for assessing school-wide positive behavior support. *Journal of Positive Behavior Interventions, 6*, 3-12.

Hsu, C. & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation, 12,* 1-8.

IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Irwin, 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*, 131-147.

Kartub, D. T., Taylor-Greene, S., March, R. E., & Horner, R. H. (2000). Reducing hallways noise: A systems approach. *Journal of Positive Behavior Interventions, 2,* 179-182.

Kincaid, D., Childs, K., Blase, K. A., & Wallace, F. (2007). Identifying barriers and facilitators in implementing schoolwide positive behavior support. *Journal of Positive Behavior Interventions, 9,* 174-184.

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

Kincaid, D., Childs, K., & George, H. (2010). *School-wide benchmarks of quality, Revised.* Unpublished instrument, University of South Florida.

Kratochwill, T. R., & Shernoff, E. S. (2004). Evidence-based practice: Promoting evidence-based interventions in school psychology. *School Psychology Review, 33,* 34-48.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from: http://quantpsy.org.

Lewis, T. J., Powers, L. J., Kelk, M. J., & Newcomer, L. L. (2002). Reducing problem behaviors on the playground: An investigation of the application of schoolwide positive behavior supports. *Psychology in the Schools, 39,* 181-190.

Lewis, T. J., & Sugai, G. (1999). Effective behavior support: A systems approach to proactive schoolwide management. *Focus on Exceptional Children, 31,* 1-24.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1-55.

Lord, F. M. (1953). On the statistical treatment of football numbers. *The American Psychologist, 8,* 750-751.

Marchant, M., Heath, M. A., & Miramontes, N. Y. (2012). Merging empiricism and humanism: Role of social validity in the school-wide positive behavior support model. *Journal of Positive Behavior Interventions, 15,* 221-230.

May, S., Ard, W., Todd, A. W., Horner, R. H., Glasgow, A., Sugai, G., et al. (2006). *School-wide information system (SWIS)*. Eugene: Educational and Community Supports, University of Oregon.

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

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

National Education Association. (2007). *Truth in labeling: Disproportionality in special education.* Washington, D.C.: Author.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education, 15,* 625-632.

O'Neill, R. E., Horner, R. H., Albin, R. W., Sprague, J. R., Storey, K., & Newton, J. (1997). *Functional assessment and program development for problem behavior: A practical handbook* (2nd ed.). Pacific Grove, California: Brooks/Cole.

OSEP Technical Assistance Center on Positive Behavioral Interventions & Supports. (2009a). *Is school-wide positive behavior support an evidence-based practice?* Retrieved from: http://www.pbis.org/ common/pbisresources/publications/EvidenceBaseSWPBS08_04_08.doc

OSEP Technical Assistance Center on Positive Behavioral Interventions & Supports. (2009b). *What is school-wide positive behavioral interventions & supports?* Retrieved from: http://www.pbis.org/common/cms/documents/WhatIsPBIS/ WhatIsSWPBS.pdf.

OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. (2013a). Retrieved from: http://www.pbis.org.

OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. (2013b). *Positive behavioral supports and the law*. Retrieved from http://www.pbis.org/school/pbis_and_the_law/default.aspx.

Pauken, P. D., & Daniel, P. T. K. (1999). Race and disability discrimination in school discipline: A legal and statistical analysis. *Education Law Reporter, December 1999.*

Ponti, C. R., Zins, J. E., & Graden, J. L. (1988). Implementing a consultation-based service delivery system to decrease referrals for special education: A case study of organization al considerations. *School Psychology Review, 17,* 89-100.

Putnam, R. F., Handler, M. W., Ramirez-Platt, C. M., & Luiselli, J. K. (2003). Improving student bus-riding behavior through a whole-school intervention. *Journal of Applied Behavior Analysis, 36,* 583-590.

Reschly, D. J. (2004). Paradigm shift, outcomes criteria, and behavioral interventions: Foundations for the future of school psychology. *School Psychology Review, 33,* 408-416.

Reynolds, C. R., & Kamphaus, R. W. (200*6). BASC-2: Behavior Assessment System for Children, Second Edition.* Upper Saddle River, NJ: Pearson Education, Inc.

Ross, S. W., Romer, N., & Horner, R. H. (2012). Teacher well-being and the implementation of school-wide positive behavior interventions and supports. *Journal of Positive Behavior Interventions, 14,* 118-128.

Scott, T. M., & Barrett, S. B. (2004). Using staff and student time engaged in disciplinary procedures to evaluate the impact of school-wide PBS. *Journal of Positive Beahvior Interventions, 6,* 21-27.

Shapiro, E. S. (2000). School psychology from an instructional perspective: Solving big, not little problems. *School Psychology Review, 29,* 560-572.

Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review, 38,* 476 – 495.

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

Solomon, B. G., Klein, S. A., Hintze, J. M., Cressey, J. M., & Peller, S. L. (2012). A meta-analysis of school-wide positive behavior support: An exploratory study using single-case synthesis. *Psychology in the Schools, 49,* 105-121.

Spaulding, S. A., Horner, R. H., May, S. L., & Vincent, C. G. (2008, November). *Evaluation brief: Implementation of school-wide PBS across the United States*. OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. Retrieved from: http://pbis.org/evaluation/evaluation_briefs/default.aspx.

Sprague, J.R., Vincent, C.G., Tobin, T.J., & CHiXapkaid (D. Pavel) (2012) Preventing disciplinary exclusions of students from American Indian/Alaska Native backgrounds. In J.S. Kaye, K.D. Cataldo, & T.A. Lang (Eds). Keeping kids in school and out of courts: A collection of reports to inform the National Leadership Summit on School Justice Partnerships. Albany, NY: New York State Permanent Judicial Commission on Justice for Children (83-95).

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Stokes, T. F. & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10,* 349-367.

Sugai, G., Horner, R. H., Algozzine, R., Barrett, S., Lewis, T.,…Simonsen, B. (2010). *School-wide positive behavior support: Implementers' blueprint and self-assessment*. Eugene, OR: University of Oregon.

Sugai, G., Horner, R. H., Dunlap, G., Hieneman, M., Lewis, T. J., Nelson, C. M., et al. (2000). Applying positive behavior support and functional behavioral assessment in schools. *Journal of Positive Behavior Interventions, 2,* 131-143.

Sugai, G., Horner, R. H., & Lewis-Palmer, T. (2001). *Team implementation checklist.* Eugene, OR: Educational & Community Supports, University of Oregon.

Sugai, G., Horner, R. H., Lewis-Palmer, T., & Rossetto Dickey, C. (2011). *Team implementation checklist, Version 3.1.* Eugene, OR: Educational & Community Supports, University of Oregon.

Sugai, G., Horner, R., & McIntosh, K. (2007). Best practices in developing a broad-scale system of school-wide positive behavior support. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V.* Bethesda, MD: National Association of School Psychologists.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. H. (2001). *The School-Wide Evaluation Tool (SET).* Eugene, Oregon: University of Oregon.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. H. (2005). *School-wide evaluation tool, Version 2.1.* Eugene, OR: Education and Community Supports, University of Oregon.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. *Journal of Emotional and Behavioral Disorders, 8,* 94 - 101.

Sugai, G., Todd, A. W., & Horner, R. H. (2000). *Effective Behavior Support (EBS) Survey: Assessing and planning behavior supports in schools.* Eugene: University of Oregon.

SurveyGizmo [Web-based Software]. (2014). Boulder, CO: www.surveygizmo.com.

Tobin, T. J. (2006). *Use of the Team Implementation Checklist in regular and alternative high schools.* Eugene, Oregon: Educational and Community Supports, University of Oregon.

Tobin, T. J., & Sugai, G. M. (1999). Using sixth-grade school records to predict school violence, chronic discipline problems, and high school outcomes. *Journal of Emotional and Behavioral Disorders, 7,* 40-53.

Tobin, T. J., & Vincent, C. G. (2011). Strategies for preventing disproportionate exclusions of African American students. *Preventing School Failure, 55,* 192-201.

Tobin, T. J., Vincent, C. G., Horner, R. H., Rossetto Dickey, C., & May, S. A. (2012). Fidelity measures used by schools to improve implementation of positive behavior interventions and supports. Manuscript in preparation.

Todd, A. W., Lewis-Palmer, T., Horner, R. H., Sugai, G., & Phillips, D. (2002). *A guide to understanding and using the SET.* Eugene, Oregon: University of Oregon.

Todd, A. W., Lewis-Palmer, T., Horner, R. H., Sugai, G., Sampson, N. K., & Phillips, D. (2004). *School-wide evaluation tool implementation manual.* Eugene, Oregon: University of Oregon.

Turnbull, A., Edmonson, H., Giggs, P., Wickman, D., Sailor, W., Freeman, R.,…Warren, J. (2002). A blueprint for schoolwide positive behavior support: Implementation of three components. *Exceptional Children, 68,* 377-402.

U.S. Department of Education, Office of Special Education and Rehabilitative Services, Office of Special Education Programs. (2011). *30th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2008*. Washington, D.C. Retrieved from: http://www2.ed.gov/about/reports/annual/osep/index.html.

Vincent, C., Spaulding, S., & Tobin, T. J. (2010). A reexamination of the psychometric properties of the School-Wide Evaluation Tool (SET). *Journal of Positive Behavior Interventions, 12,* 161-179.

Vincent, C. G., & Tobin, T. J. (2011). The relationship between implementation of school-wide positive behavior support (SWPBS) and disciplinary exclusion of students from various ethnic backgrounds with and without disabilities. *Journal of Emotional and Behavioral Disorders, 19*, 217-232.

Vincent, C. G., & Tobin, T. J. (2012). How to measure school-wide positive behavioral interventions and supports implementation fidelity with the Team Implementation Checklist: Percent of points or percent of items fully implemented. Evaluation brief. Eugene, OR: Educational and Community Supports, University of Oregon.

Walker, H. M., Horner, R. H., Sugai, G., & Bullis, M. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. Journal of Emotional and Behavioral Disorders, 4, 194 – 209.

Wang, M. D., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Wolf, M. M. (1978). Social validity: the case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203-214.

APPENDIX A

SCHOOL-WIDE EVALUATION TOOL (SET; VERSION 2.1) SCORING GUIDE

For the SET Version 2.1 please refer to Sugai et al., 2005.

APPENDIX B

INVITATION EMAIL TO SWPBIS EXPERTS TO PARTICIPATE

Dear ParticipantName,

My name is Alison Bloomfield and I am a doctoral candidate in the School Psychology program at Temple University. I am writing to you based on the number of peer-reviewed journal publications you have contributed in the area of School-wide Positive Behavioral Interventions and Supports (SWPBIS). I hope that you might be willing to participate in a brief survey on SWPBIS for my dissertation research.

The purpose of the research is to gain expert opinions on the relative importance of items on the School-wide Evaluation Tool (SET) to successful SWPBIS implementation. Ratings gained from this survey will be used to develop a new method of scoring the SET and compare the predictive validity of the new reweighted scoring and the original SET scoring in relation to a second measure of SWPBS implementation fidelity. As an accomplished SWPBIS practitioner and researcher, your expertise and experiences with SWPBIS implementation are invaluable in re-examining the factors and components of tiered behavioral support systems which are most critical to sustainable implementation.

If you would like to participate, please follow the link below to learn more about the study and access the survey. The estimated duration of your participation is less than twenty minutes.

Survey Link: http://s-746b82-i.sgizmo.com/s3/i-0000000-665825/

Please feel free to contact us if you have any questions about the research. Thank you for taking the time to help us!

Sincerely,

Alison E. Bloomfield, M.Ed., NCSP
Doctoral Candidate
School Psychology Program
Temple University
College of Education - Psychological, Organizational, and Leadership Studies
Email: alison.bloomfield@temple.edu

Catherine A. Fiorello, Ph.D., NCSP, ABPP
Professor and Program Coordinator
Temple University
College of Education – Psychological, Organizational, and Leadership Studies
School Psychology Program
Email: catherine.fiorello@temple.edu

APPENDIX C

EXPERT SURVEY INFORMED CONSENT

An Investigation of the Content and Concurrent Validity
of the School-wide Evaluation Tool

Alison E. Bloomfield, Doctoral Candidate

Catherine A. Fiorello, Professor, Program Director

Temple University College of Education - Psychological, Organizational, and Leadership Studies, School Psychology Program

This study involves research. The purpose of the research is to gain expert opinions on the relative importance of items on the School-wide Evaluation Tool (SET) to successful School-Wide Positive Behavioral Interventions and Support (SW-PBIS) implementation. Ratings gained from this survey will be used to develop a new method of scoring the SET and compare the predictive validity of the new reweighted scoring and the original SET scoring in relation to a second measure of SWPBIS implementation fidelity.

What you should know about a research study:

- This cover letter will explain this research study to you.
- You volunteer to be in a research study.
- Whether you take part is up to you.
- You can choose not to take part in the research study.
- You can agree to take part now and later change your mind.
- Whatever you decide, it will not be held against you.
- Feel free to ask all the questions you want before and after you decide.
- By clicking on the link below, you are not waiving any of the legal rights that you otherwise would have as a participant in a research study.

The estimated duration of your study participation is twenty minutes.

The study procedures consist of asking you, as an expert, to participate in an electronic survey. If you agree to participate, you will be asked to complete one or two surveys over the course of the next three months. Your survey data will be kept confidential and will only be reported in the aggregate. As a follow-up, you may be contacted and asked to re-rate or provide comments on some items. We anticipate one to two rounds of ratings.

The reasonably foreseeable risks or discomforts of your participation in this study are the donation of your time, although you may complete the survey at a time of your choosing.

The benefit you will obtain from the research is knowing that you have contributed to the understanding of this topic. At the conclusion of the study, we will send you a summary of the results.

The alternative to participating is not to participate.

There are no known risks in completing this brief survey and your responses are anonymous due to its administration through SurveyGizmo.

Please contact the research team with questions, concerns, or complaints about the research and any research-related injuries by contacting Alison Bloomfield via email: Alison.Bloomfield@temple.edu.

This research has been reviewed and approved by the Temple University Institutional Review Board (Protocol Number: 22247). Please contact them at (215) 707-3390 or e-mail them at: irb@temple.edu for any of the following: questions, concerns, or complaints about the research; questions about your rights; to obtain information; or to offer input.

Confidentiality: Efforts will be made to limit the disclosure of your personal information, including research study records, to people who have a need to review this information. However, the study team cannot promise complete secrecy. For example, although the study team has put in safeguards to protect your information, there is always a potential risk of loss of confidentiality. There are several organizations that may inspect and copy your information to make sure that the study team is following the rules and regulations regarding research and the protection of human subjects. These organizations include the IRB, Temple University, its affiliates and agents, Temple University Health System, Inc., its affiliates and agents, the study sponsor and its agents, and the Office for Human Research Protections.

We hope that you will take a few moments to complete this survey at the link below to SurveyGizmo. Please feel free to answer as many or as few questions as you wish; however, we hope that you will answer them all! If you would like to participate in our study, please click on the "YES" link below to complete the survey about the relative importance of items on the SET.

Completing the survey implies your consent to participate in the study. If you choose not to participate, simply disregard this email or exit out of the survey browser.

Please feel free to contact us if you have any questions about the research. Thank you for taking the time to help us!

Sincerely,

Alison E. Bloomfield, M.Ed.
Doctoral Candidate

School Psychology Program
Temple University – Psychological, Organizational, and Leadership Studies
Alison.Bloomfield@temple.edu

Catherine A. Fiorello, Ph.D., NCSP, ABPP
Professor and Program Coordinator
Temple University
College of Education – Psychological, Organizational, and Leadership Studies
School Psychology Program
catherine.fiorello@temple.edu

(CLICK "YES" TO CONSENT AND BEGIN THE SURVEY)

APPENDIX D

SET VALIDATION SURVEY

The following survey is aimed at evaluating the relative importance and weight of factors measured by the School-wide Evaluation Tool (SET; Sugai, Lewis-Palmer, Todd, & Horner, 2001; Version 2.1, 2005). The goal of the study is to look at whether re-weighting items on the SET to match the ratings of experts in the field of school-wide positive behavior intervention and support (SWPBIS) might provide a more valid and therefore more useful measure of SWPBIS implementation fidelity. Your ratings will be used to develop mean ratings of the relative importance of each item, which will then be used to develop a new scoring algorithm for the SET. Though only one round of surveying is expected, if there are large discrepancies in participants' responses, there may be a follow-up round to seek consensus.

Please answer the following demographic questions and survey items:

Demographic Information:

1. What is your highest level of education?
2. Approximately how many years have you been studying school-wide positive behavior interventions and supports?
3. What is your occupation?
   - Professor
   - Researcher at Independent SWPBIS Implementation Agency
   - Researcher at State SWPBIS Implementation Agency
   - Graduate Student
   - School-based Practitioner
4. What is your gender?
   - Male
   - Female
5. What is your ethnicity?
   - Hispanic or Latino
   - Not Hispanic or Latino
6. What is your race? Select one or more.
   - American Indian or Alaska Native
   - Asian
   - Black or African American
   - Native Hawaiian or Other Pacific Islander
   - White
   - Other

Survey Questions:

1. Please rate the relative importance of each of the following factors/components in the development of a sustainable implementation of a SWPBIS program.

Item	Rating					Comments (optional)
	1 Not at All Important for Sustainable Implementation of SWPBIS	2 Helpful but not Necessary for Sustainable Implementation of SWPBIS	3 Important, but not Necessary for Sustainable Implementation of SWPBIS	4 Very Important, but not Critical for Sustainable Implementation of SWPBIS	5 Critical for Sustainable Implementation of SWPBIS	
Is there documentation that staff has agreed to 5 or fewer positively stated school rules/ behavioral expectations?	1	2	3	4	5	
Are the agreed upon rules & expectations publicly posted in 8 of 10 locations? (See interview & observation form for selection of locations).	1	2	3	4	5	
Is there a documented system for teaching behavioral expectations to students on an annual basis?	1	2	3	4	5	
Do 90% of the staff asked state that teaching of behavioral expectations to students has occurred this year?	1	2	3	4	5	
Do 90% of team members asked state that the school-wide program has been taught/reviewed with staff on an annual basis?	1	2	3	4	5	
Can at least 70% of 15 or more students state 67% of the school rules?	1	2	3	4	5	
Can 90% or more of the staff asked list 67% of the school rules?	1	2	3	4	5	
Is there a documented system for rewarding student behavior?	1	2	3	4	5	
Do 50% or more students asked indicate they have received a reward (other than verbal praise) for expected behaviors over the past two months?	1	2	3	4	5	
Do 90% of staff asked indicate they have delivered a reward (other than verbal praise) to students for expected behavior over the past two months?	1	2	3	4	5	
Is there a documented system for dealing with and reporting specific behavioral violations?	1	2	3	4	5	

Do 90% of staff asked agree with administration on what problems are office-managed and what problems are classroom–managed?	1	2	3	4	5	
Is the documented crisis plan for responding to extreme dangerous situations readily available in 6 of 7 locations?	1	2	3	4	5	
Do 90% of staff asked agree with administration on the procedure for handling extreme emergencies (stranger in building with a weapon)?	1	2	3	4	5	
Does the discipline referral form list (a) student/grade, (b) date, (c) time, (d) referring staff, (e) problem behavior, (f) location, (g) persons involved, (h) probable motivation, & (i) administrative decision?	1	2	3	4	5	
Can the administrator clearly define a system for collecting & summarizing discipline referrals (computer software, data entry time)?	1	2	3	4	5	
Does the administrator report that the team provides discipline data summary reports to the staff at least three times/year?	1	2	3	4	5	
Do 90% of team members asked report that discipline data is used for making decisions in designing, implementing, and revising school-wide effective behavior support efforts?	1	2	3	4	5	
Does the school improvement plan list improving behavior support systems as one of the top 3 school improvement plan goals?	1	2	3	4	5	
Can 90% of staff asked report that there is a school-wide team established to address behavior support systems in the school?	1	2	3	4	5	
Does the administrator report that team membership includes representation of all staff?	1	2	3	4	5	
Can 90% of team members asked identify the team leader?	1	2	3	4	5	
Is the administrator an active member of the school-wide behavior support team?	1	2	3	4	5	
Does the administrator report that team meetings occur at least monthly?	1	2	3	4	5	
Does the administrator report that the team reports progress to the staff at least four times per year?	1	2	3	4	5	
Does the team have an action plan with specific goals that is less than one year old?	1	2	3	4	5	
Does the school budget contain an allocated amount of money for building and maintaining school-wide behavioral support?	1	2	3	4	5	
Can the administrator identify an out-of-school liaison in the district or state?	1	2	3	4	5	

2. What SWPBIS implementation factors or components do you believe are missing from the SET? Would you add any items?

3. What tool do you prefer to use for SWPBIS implementation evaluation? Why?

APPENDIX E

TEAM IMPLEMENTATION CHECKLIST (VERSION 3.1)

For the TIC Version 3.1 please refer to Sugai et al., 2011.

APPENDIX F

SET VALIDATION SURVEY EXPERTS' COMMENTS ON

MISSING FACTORS FROM THE SET

Frequency of Comment	Missing Factor
4	Classroom Systems
3	Culturally Responsive Implementation/ Cultural Relevance of Behavior Support
2	Family Engagement/Involvement
2	Break Down Elements of SWPBIS Using Task Analysis • Make it more difficult to obtain high scores • Allow for more specific action planning
2	Include Secondary and Tertiary Tiers
1	Bullying Prevention
1	Direct Observation of Staff
1	Emphasize Quality Indicators of Key Elements – Not Just Presence
1	Systematic Supervision of Common Areas
1	Faculty Involvement and Buy-In
1	Emphasis on Incorporating/Integrating Teaching Rules and Expectations into Curriculum Throughout Year

Note: Expert comments on factors/elements missing from the SET have been paraphrased and summarized.

APPENDIX G

SET VALIDATION SURVEY EXPERTS' PREFERRED TOOLS FOR ASSESSING SWPBIS IMPLEMENTATION

Frequency	Preferred Tool	Comments
9	School-wide Evaluation Tool (SET)	• Missing direct observation • Only addresses Tier 1 supports • Only when used by an objective, external observer-interviewer • Time intensive • Objective/comprehensive measure of Tier I • Relatively efficient • Research and evaluation tool
7	Benchmarks of Quality (BOQ)	• Easier to score than SET • Very detailed • All stages of implementation and sustainability • More comprehensive than SET • Requires fewer resources (personnel/time) • Useful for action planning • More relevant than SET • Validated and accurate
3	Team Implementation Checklist (TIC)	• More detail for ongoing progress monitoring and improvement • Helpful guide for new teams
2	PBIS Tiered Fidelity Inventory	• Assesses all 3 tiers • Draws from multiple instruments • Can be used summatively for annual assessment • Tier assessments can be used to progress monitor
2	Effective Behavior Survey / Self Assessment Survey (EBSSAS)	• Introduces common SWPBIS language in schools beginning implementation
1	Direct Observation in Classroom	• Assesses SWPBIS practices in the classroom
1	Individual Student Systems Evaluation Tool (ISSET)	• Assess all 3 tiers
1	Implementation Phases Inventory (IPI)	• Helpful during first years of implementation
1	PBS Implementation Checklist (PIC)	• Validated; Good for progress monitoring
1	Benchmarks for Advanced Tiers (BAT)	• Good for assessing Tier 3

Note: Expert comments on preferred methods of assessing SWPBIS implementation fidelity have been paraphrased and summarized. Some experts reported more than one preferred tool or specified circumstance under which it is preferred.

APPENDIX H

SET VALIDATION SURVEY RAW EXPERT RATINGS OF SET ITEMS

	Participant ID																			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
SET Item A1	5	5	5	5	4	4	5	5	5	3	5	4	5	5	5	5	5	5	5	5
SET Item A2	5	4	4	4	4	5	5	5	4	5	3	3	5	5	5	5	5	5	3	4
SET Item B1	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5
SET Item B2	5	5	4	5	5	3	5	5	5	4	2	4	5	2	5	5	5	5	4	5
SET Item B3	4	5	3	5	5	3	5	5	5	4	2	4	5	2	5	5	5	5	4	5
SET Item B4	5	4	3	5	5	4	5	5	5	4	4	5	5	5	5	5	4	5	5	5
SET Item B5	5	5	4	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5
SET Item C1	5	4	4	3	4	4	4	5	5	5	5	5	5	4	4	5	5	5	5	5
SET Item C2	5	4	5	3	5	3	5	5	5	3	4	4	5	5	4	5	4	4	4	4
SET Item C3	5	5	5	3	5	3	5	5	5	5	5	4	5	5	4	5	4	4	3	4
SET Item D1	5	5	4	5	4	5	5	5	5	5	4	4	4	4	5	5	5	5	5	4
SET Item D2	5	5	3	5	5	4	5	5	5	5	5	4	4	3	5	5	5	5	4	5
SET Item D3	3	4	3	4	2	5	5	1	5	4	2	3	4	3	3	2	3	5	2	2
SET Item D4	3	5	3	5	2	5	5	1	5	3	4	4	4	3	3	2	3	5	2	3
SET Item E1	4	5	3	5	4	4	5	5	5	4	3	4	5	5	5	2	5	5	5	5
SET Item E2	5	5	4	4	5	4	5	5	5	3	2	4	5	5	5	5	5	5	5	5
SET Item E3	5	5	3	3	5	3	5	5	5	5	2	4	5	5	5	3	5	4	3	5
SET Item E4	5	5	5	5	5	5	5	5	5	5	3	4	5	5	5	3	5	4	5	4
SET Item F1	4	4	4	3	4	5	5	5	4	5	2	4	4	4	3	2	3	5	3	3
SET Item F2	5	5	4	4	4	5	5	5	5	4	3	4	5	5	5	4	4	4	3	3
SET Item F3	5	4	4	3	4	3	5	5	5	2	3	3	5	3	5	3	4	4	3	3
SET Item F4	4	4	4	3	4	3	5	5	4	2	3	4	5	2	5	3	4	4	4	3
SET Item F5	5	5	5	5	5	4	5	5	5	5	3	4	4	5	5	5	5	5	5	5
SET Item F6	5	5	5	3	5	3	5	5	4	4	2	3	4	5	5	5	4	5	5	5
SET Item F7	4	5	2	4	5	4	4	5	5	5	3	3	4	5	5	2	4	4	4	4
SET Item F8	5	5	3	3	5	5	4	5	5	5	4	4	4	5	5	4	4	5	5	5
SET Item G1	4	4	3	3	3	3	4	5	5	5	4	4	5	4	5	2	4	5	3	3
SET Item G2	4	5	3	3	3	2	4	5	5	2	4	3	5	3	4	2	3	4	3	4