MEASURING PRIMARY PLAN TREATMENT INTEGRITY OF COMPREHENSIVE, INTEGRATED THREE-TIERED PREVENTION MODELS

By

Allison Leigh Bruhn

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Special Education

August, 2011

Nashville, Tennessee

Approved:

Professor Kathleen Lane, Chair

Professor Vicki Harris

Professor Kim Paulsen

Professor Joe Wehby

TABLE OF CONTENTS

To my late softball coach, Gary Page, who taught me perseverance and perspective.

To my daughter, Loxton Paige, my heart and soul.

ACKNOWLEDGEMENTS

In four years at Vanderbilt, I have been blessed to have the mentorship, support, and friendship of many people. First, I am so grateful for Dr. Kathleen Lane for bringing me from Fairview Middle School to Peabody to study under her advisement. She pushed me beyond what I thought possible, provided opportunities for success, and encouraged me when I needed it most. Additionally, I would like to thank my committee members Dr. Kim Paulsen, Dr. Vicki Harris, and Dr. Joe Wehby for asking questions, being flexible, and being genuinely supportive of my work and my life. I am also extremely grateful for the work of the PSI staff, as well as Dr. Wendy Oakes, whose attention to detail proved invaluable. Further, I literally owe my life to Dr. Ted Hasselbring. Without him, I would not have lived to see this project to completion.

While I professionally and personally appreciate each member of the "Elite 11", I am particularly grateful to my friend, colleague, and fellow "Navy Seal," Mary Crnobori. Her support and loyalty were unparalleled. Another friend and colleague for whom I am eternally grateful is Shanna Eisner Hirsch, whose work ethic was inspiring and friendship unwavering.

Finally, I am indebted to my family near and far. My parents, in-laws, sister, daughter, and husband offered so much love and support, but more importantly, patience and understanding when I worked on holidays, weekends, and vacations. I am humbled by my husband Michael's selflessness and love as he was always willing to give me the time and space I needed to get things done. I cannot thank all of them enough for allowing me to pursue my dreams.

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

For years, the terms “reliability”, “integrity”, and “fidelity” have been used interchangeably when describing the quality and accuracy of intervention procedures (Wolery & Ledford, 2011). Although the aforementioned terms are synonymous, they hold different meaning when preceded by the words “procedural” or “treatment” (Wolery & Ledford). Procedural fidelity (reliability, integrity), the gold standard for assessing study procedures, is a broad measure of all relevant components in all conditions (Billingsley, White, & Munson, 1980; Gast, 2010). Namely, procedural fidelity involves measuring all procedures in baseline (or control) *and* intervention conditions (Wolery & Ledford). Treatment integrity (reliability, fidelity), on the other hand, is a narrower measure of the degree to which the independent variable is implemented as intended (Yeaton & Sechrest, 1981). To clarify, researchers measuring treatment integrity are concerned with implementation of the independent variable during the treatment condition only, while those measuring procedural fidelity examine treatment integrity during intervention but also the extent to which the independent variable and other procedural variables were used during baseline or control conditions (Wolery & Ledford). By assessing procedural fidelity, researchers can better understand how pre-intervention (i.e., baseline or control) and intervention conditions differ in relation to the independent variable (Wolery & Ledford). Additionally, measuring procedural fidelity provides a more extensive assessment of procedural variables in *all* conditions and allows

researchers to (a) detect threats to internal validity and (b) be more precise in interpreting findings, most research has focused simply on reporting treatment integrity (Wolery & Ledford).

While it is ideal to measure procedural variables in both baseline (or control) and treatment conditions, measuring treatment integrity of the intervention is a fundamental starting point to understanding the implementation of procedural variables. Researchers agree it is imperative to understand the quality and accuracy of treatment implementation (i.e., treatment integrity) before drawing conclusions about an intervention's effectiveness (Yeaton & Sechrest, 1981). For example, if significant changes in behavior occur, but no data are presented about implementation levels then precise conclusions regarding intervention effectiveness cannot be drawn. In a related way, if no changes in behavior occur and no integrity data are presented, we cannot accurately distinguish between an ineffective intervention and a poorly executed, but effective intervention (Gresham, Gansle, & Noell, 1993; Yeaton & Sechrest). In sum, we recognize procedural fidelity measures provide more detailed information about a study. In this study, however, we focus on treatment integrity as it is a good first and essential step to understanding intervention effectiveness.

Further, our field has just recently begun to recognize the importance of treatment integrity, let alone procedural fidelity, as demonstrated in the literature. Multiple reviews of educational and psychological research have revealed a dearth of treatment integrity data reported in intervention studies (e.g., Durlak & Dupre, 2008; Gresham et al., 1993; Lane, Kalberg, & Edwards, 2008; Lane, Robertson, & Graham-Bailey, 2006; McIntyre, Gresham, DiGennaro, & Reed, 2007; Moncher & Prinz, 1991). Given the research

community's assertion treatment integrity is necessary for drawing accurate conclusions about intervention effectiveness (Gersten et al., 2005; Gresham et al.; Horner et al., 2005), the findings of these reviews are disconcerting. As the field moves towards (a) establishing evidence-based practices (EBPs), and (b) determining how resources are allocated in three-tiered prevention models which hinge on employing EBPs, treatment integrity (and eventually, procedural fidelity) must be central to the discussion (Schulte, Easton, & Parker, 2009). Thus, the instructional and administrative decision-making process used in three-tiered prevention models should include, at a minimum, measurement of treatment integrity.

Three-tiered Prevention Models

Several types of three-tiered models of prevention exist, including: (a) those used to place students in special education services for learning disabilities (e.g., response-to-intervention, RtI); (b) schoolwide positive behavior support (SWPBS) programs focusing solely on behavior; (c) and comprehensive, integrated three-tiered (CI3T) models including academic, behavioral, and social components (Lane, Menzies, Oakes, & Kalberg, 2011). The underlying structure for supporting students in all types of three-tiered prevention models (e.g., RtI, SWPBS, CI3T) is the same. Specifically, each model offers a continuum of support that increases in intensity in response to student need with the goal of early student identification and intervention. This requires student progress to be monitored continuously, which is fundamental to making data-based decisions that inform instruction and guide student placement in various intervention levels (Batsche et al., 2005).

Primary level. At the base, or primary, level; all students are exposed to prevention efforts. In RtI models, the primary level may include a core academic curriculum (e.g., evidence-based reading instruction; Compton, D. Fuchs, L. S. Fuchs, & Bryant, 2006; Martson, 2005) and universal screening to assess academic performance in various content areas. These benchmark assessments (e.g., DIBELS; Kaminski & Good, 1996), which may occur three times per year, are used to determine which students may go on to receive tier two, or secondary, supports. In SWPBS models, the primary prevention level consists of modeling, teaching, and reinforcing three to five schoolwide expectations. Similar to RtI models, progress is monitored according to goals of the primary plan. Schoolwide behavioral data such as office discipline referrals (ODRs) or systematic behavior screeners can be used as an index of responsiveness (e.g., Lane, Kalberg, Bruhn, Mahoney, & Driscoll, 2008; Sugai, Sprague, Horner, & Walker, 2006). In CI3T models, schools employ evidence-based, core curricula; the SWPBS framework for establishing behavioral expectations; as well as a social component which may include teaching all students a validated social skills curriculum (e.g., *Social Skills Improvement System: Intervention Guide*, SSIS; Elliott & Gresham, 2008). Access to academic and social skills curricula is facilitated by the SWPBS framework which allows students to experience consistent expectations across settings. To determine responsiveness and make decisions regarding future support, academic and behavioral data are analyzed in tandem (Lane, Menzies, & Kalberg, in press). Generally, it is expected that about 80% of the student population will make adequate progress with exposure to only the primary level of prevention (Batsche et al., 2005; Sugai & Horner, 2006).

Secondary level. Students who are non-responsive to the primary plan may go on to receive secondary, or tier two, supports. It is estimated 10-15% of the school population will need this level of support (Batsche et al., 2005; Sugai & Horner, 2006). Secondary supports, while less costly in terms of time and resources than tertiary supports, are often targeted at small groups for instruction on academic, behavioral, and/or social skills. Or, they may constitute general supports for multiple individuals. Examples include specific literacy training (e.g., Wanzek & Vaughn, 2008), self-regulated strategy development for writing (e.g., Lane, Graham, et al., 2009), check-in/check-out (CICO) procedures (e.g., Fairbanks, Sugai, Guardino, & Lathrop, 2007), study skills instruction (e.g., Robertson & Lane, 2007) and explicit social skills instruction (e.g., Gresham, Van, & Cook, 2006). As with primary prevention, progress monitoring to inform instruction and guide placement also occurs at the secondary level. Additionally, similar to the importance of understanding treatment integrity levels of the primary plan before moving students into secondary supports, it is equally as important to understand secondary intervention treatment integrity levels prior to placing a student in a tertiary intervention.

Tertiary level. The tertiary level of support is designed to serve about 5% of the school population, as it is reserved for the students exhibiting the most need for intervention (Batsche et al., 2005; Sugai & Horner, 2006). Students requiring tertiary support often have been non-responsive to primary or secondary efforts, and thus, require an intervention that is more individualized than those offered at the primary and secondary levels. In academic-only RtI models, tertiary support is used to address specific learning deficits (Hawken, Vincent, & Schumann, 2008). From a behavioral

perspective, functional assessment-based interventions (FABIs) often constitute tertiary support. Further, in any three-tiered model, tertiary support may occur across multiple settings beyond the school (e.g., home, community) and involve multiple stakeholders (e.g., parents, teachers, school psychologist). Because interventions are highly individualized to meet the specific academic, behavioral, and/or social needs of the student, tertiary support is more time and resource-intensive than other support levels. Consequently, it is imperative we first determine how less-intense support levels (i.e., primary and secondary) are being implemented prior to placing students in costly tertiary interventions that may not be warranted.

To summarize, in all three-tiered prevention models (e.g., RtI, SWPBS, CI3T) decision-makers must know if the primary prevention level was implemented as planned prior to placing a student in a targeted intervention (Bruhn, Lane, & Hirsch; 2011). Schools with a paucity of resources (e.g., time, materials, personnel) must exercise discretion when allocating those resources for targeted or intensive supports, which are generally costly. Essentially, schools must determine if secondary or tertiary supports are warranted based on a student's response to less-intense levels of intervention. For example, if failure to respond to instruction provides the basis for a special education placement in a RtI model, it is important to understand teacher adherence to curriculum and quality of instructional delivery, as well as student exposure and responsiveness (Schulte et al., 2009). This decision begins with measuring primary plan treatment integrity and then deciding if the primary plan needs to be implemented with greater accuracy or if the plan is being implemented accordingly and, therefore, a student can be identified as non-responsive and in need of more support (Bruhn et al.). By

understanding the extent to which students are exposed to the primary program (i.e., treatment integrity), school personnel can make quality, data-based decisions about student responsiveness.

Treatment Integrity at the Primary Prevention Level

To date, the majority of articles on SWPBS outcomes have used the Schoolwide Evaluation Tool (SET; Sugai et al., 2001) to measure primary-level implementation using the whole school as the unit of analysis. The SET, which is conducted annually or semi-annually, measures treatment integrity at the school level across seven domains or subscales (i.e., Expectations Defined, Behavioral Expectations Taught, On-Going System for Rewarding Behavioral Expectations, System for Responding to Behavioral Violations, Monitoring and Decision-Making, Management, District-Level Support). First, an independent assessor interviews a school administrator, and then briefly interviews randomly selected teachers and students about the schoolwide program. Additionally, the assessor reviews permanent products (e.g., discipline handbook, office discipline referral forms) and observes the school environment. Schools are said to be implementing with high fidelity if they achieve 80/80 criteria meaning they score 80% on the Total Score and 80% in the Behavioral Expectations Taught subscale. In the original psychometric study of the SET across 17 schools, researchers found the SET internally consistent (Cronbach's alpha = .96) and demonstrated test-retest reliability at 97.3% (Horner et al., 2004). Recently, Vincent, Spaulding, and Tobin (2010) examined similarities and differences in SET data across elementary, middle, and high school levels; the internal consistency at the three school levels; and how SET scores correlated

with the Team Implementation Checklist (TIC; Sugai, Todd, & Horner, 2001; Cronbach's alpha = .93; Barrett, Bradshaw, & Lewis-Palmer, 2008), which is 17-item checklist completed by SWPBS teams either monthly or quarterly. The TIC assesses the start-up activities and ongoing monitoring procedures of SWPBS. Correlations between TIC and SET subscales ranged from .32 to .75 (elementary school), .08 to .57 (middle school), and .11 to .53 (high school). Further, the SET was more cohesive, or internally consistent, at the elementary rather than middle and high school levels (Vincent et al.). Finally, the SET may demonstrate large changes from pre-implementation to initial implementation, but as schools move toward maintenance the SET may not be sensitive enough to detect variability in implementation.

In addition to the type and level of assessment, the method of measurement is central to the discussion of treatment integrity for the primary plan. While treatment integrity of secondary and tertiary interventions is often measured through direct observation or self-report using rating scales or procedural checklists, only a handful of studies (e.g., Barrett et al., 2008; Lane, Kalberg, Bruhn, et al., 2008; Lane, Wehby, Robertson, & Rogers, 2007) have used these techniques to measure primary-level implementation. Self-assessment via a component checklist allows teachers, who are most familiar with their own behavior, to provide a more global measure of treatment integrity that may be less intrusive and less costly than direct observation (Lane, Bocian, MacMillan, & Gresham, 2004).

Recently, another evaluation tool, the Schoolwide Benchmarks of Quality (BoQ, Kincaid, Childs, & George, 2005) was developed as a self-report measure to serve as an alternative to the SET. The BoQ was developed so that schools could assess their own

fidelity on-site rather than rely on an external assessor to complete the SET (Cohen, Kincaid, & Childs, 2007). Similar to the TIC (Sugai et al., 2001), the BoQ is a 53-item rating scale across 10 domains (i.e., PBS Team, Faculty Commitment, Effective Discipline, Data Entry, Expectations and Rules, Reward System, Lesson Plans, Implementation Plans, Crisis Plans, Evaluation) completed by SWPBS leadership teams. A validation study of the BoQ (Cohen et al.) revealed it internally consistent with Cronbach's alpha equaling .96. The test-retest correlation between Time 1 and Time 2 was .94 ($p < .01$). Additionally, interrater reliability was .97 ($p < .01$). Finally, the total scores on the BoQ had a .51 correlation ($p < .05$) with total scores on the SET.

Like the TIC and BoQ, the Self-Assessment and Program Review (SAPR; Cheney & Walker, 2003a, 2003b) is a treatment integrity checklist completed by SWPBS teams to monitor implementation practices and identify goals for improvement. The SAPR contains 10 subscales with each subscale consisting of four to eight items. In a psychometric study of the SAPR (Walker, Cheney, & Stage, 2009), a survey of validity evidence based on test content revealed all items on the SAPR as mostly to fully relevant to the implementation of SWPBS. The overall alpha level was .96. And, in a descriptive comparison of schools scoring above or below 80% on subscales from both the SET and SAPR, these subscales were in agreement 71% of the time. When 75% was the criterion, agreement was 81%. Authors concluded these data provided preliminary evidence as to the concurrent validity of the SAPR and SET.

Finally, unlike the SET, TIC, BoQ, and SAPR; the Implementation Phases Inventory (IPI; Bradshaw, Barrett, & Bloom; 2004) is an assessment tool designed to identify the particular SWPBS phase a school is in, and thus help schools to set program

goals and improve implementation. Bradshaw and colleagues identified the phases as preparation, initiation, implementation, and maintenance. The IPI contains 44 questions associated with the critical features of SWPBS, routine start-up activities, program materials, and formal policies and procedures. Rather than being completed by a team, the IPI is completed by SWPBS coaches who are liaisons between the school and the district (e.g., school psychologists, guidance counselors, or school behavioral specialists). Coaches rate the 44 items using a Likert-type scale (0 = *not in place*, 1 = *partially in place*, 2 = *full implementation*). Recently, Bradshaw and colleagues (2009) found the IPI to be internally consistent (Cronbach's alpha = .94) and to produce stable results over time (i.e., test-retest reliability; r (40) = .80, $p \leq .01$). To examine interrater reliability, team leaders (in addition to coaches) at each school completed the IPI. They found moderate interrater reliability between coaches' and team leaders' scores (r = .61, $p \leq$ .01). Finally, they concluded the IPI total scores and all subscale scores were significantly ($p \leq .01$) correlated with total scores and all subscale scores of the SET and TIC.

While all of the aforementioned tools for assessing primary plan implementation of SWPBS have demonstrated some evidence of reliability and validity, none assess implementation from all or the majority of teachers in the building. Rather, scores represent the perspective of teams, a small sample of faculty, coaches, or team leaders. Further, these measures do not assess implementation at the classroom level. Ideally, schoolwide, primary plans should be implemented by every adult (e.g., custodial staff, cafeteria staff, librarians, paraprofessionals, etc.) in the school building, yet it is teachers within classrooms who most frequently provide instruction and feedback to students on

academic, behavioral, and social expectations. Because teacher behavior varies in response to student behavior in the classroom (Jack et al., 1996; Sutherland & Oswald, 2005), measuring implementation of the primary plan at the classroom level is imperative. Namely, measuring treatment integrity at the school level only and not the classroom level, does not allow for analysis and discussion of variance between classrooms and their associated implementation outcomes (Zvoch, 2009). By measuring primary plan implementation at the classroom level, teachers and administrators can make data-based decisions for placing students into more intense supports (e.g., secondary and tertiary interventions) and provide professional development opportunities or mentoring for teachers who need additional support in implementation (Bruhn et al., 2011). Thus, there is a need in the field for psychometrically sound measures that can determine the quality and accuracy of primary plan implementation within individual classrooms. To date, the field lacks classroom-level tools that reliably measure primary-plan implementation and allow valid inferences to be drawn regarding implementation and student progress.

Overview and Purpose

Prior to this study, school-site leadership teams from eight elementary schools attended a year-long training on designing, implementing, and evaluating a CI3T prevention model including a core SWPBS plan as well as academic and social components. In the years subsequent to training, faculty and staff at all eight schools elected to implement their customized CI3T plan. Implementing schools were then invited to participate in an evaluation study to assess both the treatment integrity and

social validity of the primary plan. As part of this evaluation, treatment integrity was assessed using (a) the SET, (b) 30 min direct observation of classroom teachers by research assistants (DO-RA) and teacher report of the same 30 min period (DO-T), and (c) teacher self-report of implementation since the beginning of the year (TSR, August to February).

At the time, the SET was the only treatment integrity measure included in the evaluation study with established reliability and validity evidence. Although the SET paints a solid picture of implementation for the school as a whole, it does not offer information about the extent to which the primary plan behavioral components were carried out in specific classrooms. As such, reliable and valid measures are needed to measure implementation of the primary plan in individual classrooms. Namely, as school personnel make decisions about student responsiveness to the primary plan and consider placing them into more intense supports, understanding the degree to which the primary plan is implemented in a particular student's classroom is imperative. In an effort to provide more detailed information to the schools regarding treatment integrity at the classroom level, as reported by teachers and instructional staff, evidence of the reliability and validity of the measure must be established. Further, if schools are going to use treatment integrity data to make decisions, they need to be able to rely on the accuracy of the tools used to measure treatment integrity. Therefore, the purpose of this proposed study was two-fold.

The first goal was to examine initial evidence of the reliability of the TSR by determining the internal consistency as measured by alpha coefficients. Internal consistency estimates were calculated for each subscale and the total. Because the

instrument items were selected for each domain based on (a) the research team's knowledge of CI3T models and assessment tools and (b) a previous treatment integrity evaluation study, it was expected the individual items adequately measured the constructs of interest (i.e., Procedures for Teaching, Procedures for Reinforcing, Procedures for Monitoring). Additionally, we expected TSR subscales to be moderately correlated with each other indicating they were related but not measuring the same constructs.

The second goal of this study was to determine the relationship between the SET and the TSR. We examined the correlation between each subscale on the TSR with each subscale on the SET as well as the TSR total score and SET total score. Given the SET, which is a whole-school measure, was designed to measure *only* primary plan behavioral components and the TSR, which is completed at the teacher level, was designed to measure academic, behavioral, and social components of the primary plan; it was reasonable to expect only TSR subscales with salient behavioral items to be correlated with SET subscales.

CHAPTER II

METHOD

Participants

Participants were 183 teachers and instructional staff from elementary schools in Middle Tennessee who participated in a year-long training series to design a CI3T prevention model for implementation at their respective schools. They included general educators (66.12%, n = 121), special educators (12.02%, n = 22), related service providers (1.64%, n = 3) such as school psychologists, counselors, and therapists; and other educators such as related arts teachers (e.g., art, music, physical education) and reading specialists (11.48%, n = 21) employed at these schools. The majority of participants were female (87.43%, n = 160) and held at least a Master's degree (i.e., Masters's or Master's plus 30 hours; 53.55%. n = 98). A more detailed depiction of participants by school and total sample is provided in Table 1.

Of the eight participating schools, three schools serving grades pre-kindergarten (preK) through fifth (Schools A, B, C) and one school serving preK through eighth (School H) were from District A, while four schools serving grades preK through fourth (Schools D, E, F, G) were from the adjacent District B. However, for the purposes of this study, only teachers and instructional staff from preK through fifth were included as the focus was elementary school implementation of a CI3T model. All participating schools were public schools from a variety of geographic locales including rural, town, and city

Table 1

Participant Demographics by School

Characteristic (n)	School								
	A *n* = 38	B *n* = 11	C *n* = 30	D *n* = 16	E *n* = 25	F *n* = 19	G *n* = 23	H *n* = 21	Total *n* = 183 *n* (%)
Gender									
Male	2	0	0	0	2	2	1	1	8 (4.37)
Female	34	11	30	15	14	15	21	20	160 (87.43)
Not Reported	2	0	0	1	9	2	1	0	15 (8.20)
Role									
General Education Teacher	28	6	23	10	11	13	17	13	121 (66.12)
Special Education Teacher	5	2	2	5	2	0	4	2	22 (12.02)
Related Service Provider [a]	0	1	0	0	0	1	1	0	3 (1.64)
Other[b]	2	2	5	0	3	3	0	6	21 (11.48)
Not Reported	3	0	0	1	9	2	1	0	16 (8.74)
Highest Degree Obtained									
Bachelor's	15	7	11	4	6	6	9	6	64 (34.97)
Master's	17	3	13	8	4	7	9	11	72 (39.34)
Master's +30	4	1	4	1	5	4	4	3	26 (14.21)
Other	0	0	1	1	1	0	0	0	3 (1.64)
Not reported	2	0	1	2	9	2	1	1	18 (9.84)
Years of experience *M* (*SD*)	16.61 (10.80)	13.64 (8.48)	15.93 (11.27)	6.93 (5.61)	19.57 (11.27)	11.71 (10.52)	7.84 (7.75)	16.62 (12.73)	14.01 (10.86)

Note. [a] Related service provider = school psychologist, counselor, therapist, other. [b] Other = art, reading recovery, music, physical education, reading Title I specialist

Note. *n* and % based on data received from participants (e.g., some participants chose not to disclose all demographic characteristics)

Table 2

School Demographics

Characteristic	School							
	A	B	C	D	E	F	G	H
Total number of students	715	111	477	304	252	264	405	548
Grades served	PreK-5	K-5	PreK-5	PreK-4	PreK-4	PreK-4	PreK-4	PreK-8
Ethnicity *n* (%)								
African-American	12 (1.7)	0 (0)	5 (1)	262 (86.2)	182 (72.2)	250 (94.7)	387 (95.6)	16 (2.7)
Asian/Pacific Islander	8 (1.1)	0 (0)	7 (1.5)	0 (0)	1 (0.4)	3 (1.1)	2 (.5)	5 (0.9)
Hispanic	18 (2.5)	1 (0.9)	24 (5)	2 (0.7)	25 (9.9)	3 (1.1)	11 (2.7)	17 (2.9)
Native American	2 (0.3)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.2)
White	675 (94.4)	110 (99.1)	441 (92.5)	40 (13.2)	44 (17.5)	8 (3.0)	5 (1.2)	543 (93.3)
Gender *n* (%)								
Female	359 (50.2)	48 (43.2)	228 (47.8)	130 (42.8)	125 (49.6)	125 (47.3)	189 (46.7)	268 (46.0)
Male	356 (49.8)	63 (56.8)	249 (52.2)	174 (57.2)	127 (50.4)	139 (52.7)	216 (53.3)	314 (54.0)
Economic Disadvantage Rate (%)	16.0	22.1	42.6	92.8	92.6	>95.0	>95.0	31.2
Geographic Locale	Rural: distant	Rural: distant	Town: distant	Rural: fringe	City: large	City: large	City: large	Rural: distant
Years of Implementation	4	5	3	1	1	1	1	1

Note. Data obtained from www.tn.gov (Tennessee Department of Education School Report Card) and http://nces.ed.gov/ccd (National Center for Education Statistics: Common Core of Data)

areas (see Table 2) The student demographics represented a range of ethnicities and economic disadvantage rates (see Table 2).

School Training Procedures

All eight schools participated in a year-long training series on CI3T models of prevention at Vanderbilt University and implemented the CI3T model in all years subsequent to the training. Five schools attended the training during the 2008-2009 school year, one during the 2006-2007 school year, one during the 2005-2006 school year, and one during the 2004-2005 school year. Thus, in the current study, schools had been implementing for one to five years (see Table 2). Participants at the training included school-selected members of the CI3T team. The CI3T teams consisted of, at a minimum, two general educators, one special educator, an administrator, a parent, and a student. Some teams also elected to include a counselor. Team members attended two full-day trainings and five 2-hr trainings throughout the school year. During the trainings, teams learned about the historical and legal background of CI3T models as well as a rationale for using multiple data sources (e.g., systematic behavior screeners, curriculum-based measures, treatment integrity) to assess the effectiveness of CI3T implementation. Then, the team designed a primary-level plan which included a purpose statement; a description of school expectations; procedures for teaching, reinforcing, and monitoring expectations; and the academic, behavioral, and social roles and responsibilities of students, parents, teachers, and administrators. Procedures for teaching expectations varied across schools and included procedures such as: weekly social skills lessons, teacher modeling, posters displaying expectations, teacher-led skits, and morning

announcements. Additionally, teachers and instructional staff were expected to provide engaging, differentiated instruction linked to district or state standards. Procedures for reinforcing expectations included giving students tickets paired with behavior specific praise for meeting academic, behavioral, and social expectations. Tickets were exchangeable for tangible and non-tangible items. Some schools chose to have random ticket drawings for prizes, while others used tickets for admission to assemblies and school parties. To monitor the implementation and effectiveness of the CI3T, schools selected academic, behavioral, and program measures to be assessed at specific time points throughout the year as part of regular school practices. These data were to be used to drive instructional and administrative decisions. School teams created an assessment schedule indicating what measures would be collected at specific times throughout the school year (see Figure 1). Examples of academic measures included standardized test scores, course failures, and curriculum-based measures. Behavioral measures included office discipline referrals, attendance, tardies, and systematic behavior screeners. Finally, program measures included social validity assessments (Primary Intervention Rating Scale [PIRS], Lane, Robertson, & Wehby, 2002) and treatment integrity assessments (detailed descriptions to follow). In addition to the primary-level plan, the teams also created secondary- and tertiary-level intervention grids. These grids contained descriptions of secondary and tertiary academic, behavioral, and social interventions available at the school; as well as student entry and exit criteria.

Following the development of their CI3T plans, each school solicited feedback from faculty and staff, as well as 100 randomly selected parents and students. This

Figure 1
Sample Monthly Assessment Schedule

	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
		Quarter 1		Quarter 2			Quarter 3		Quarter 4	
School Demographics										
Student Demographics		X		X			X		X	
Student Outcome Academic Measures										
STARS Reading		X		X					X	
WCS Writing		X		X			X		X	
TCAP									X	
Student Outcome Behavior Measures										
SRSS-IE		X		X					X	
Discipline: SWIS		X		X			X		X	
Attendance (Tardies/Absences Excused & Unexcused)		X		X			X		X	
Referrals										
SPED GEIT		X		X			X		X	
Program Measures *For Consented Teachers Only*										
Social Validity (PIRS)		X					X			
SET/Treatment Integrity (TI) Self-Report							X			
TI Observations									X	

Note. Adapted from *Developing schoolwide programs to prevent and manage problem behaviors: A -step-by-step approach*. By K. L. Lane, J. R. Kalberg, and H. M. Menzies. Copyright 2009 by Copyright Holder. Reprinted with permission.

feedback was used to guide changes to the plan. Once the plan was finalized, each school's faculty voted on whether or not to implement the CI3T plan in the upcoming school year. Additionally, if the plan passed the schoolwide vote, the faculty and staff at each school had the option to participate in an evaluation of the primary-level of implementation. Schools were consented to a program evaluation at the beginning of each year of CI3T implementation. This evaluation involved the collection of treatment integrity data. Namely, the degree to which the primary level of the three-tiered plan was being implemented as designed by consented faculty and staff was assessed in four different ways by Vanderbilt research assistants (RAs) and teacher self-report (see Treatment Integrity Measures for descriptions of the two measures used in the current study).

School Consenting Procedures

During the fall of 2009, a team of research assistants (RAs), the project coordinator, and/or primary investigator gave a presentation at each school. The purpose of the presentation was to inform potential participants about the risks (e.g., loss of time to complete forms) and benefits (e.g., evaluation will provide information on the effects of the plan and improve educational programming) of participating in the program evaluation, as well as the three consent levels offered. A level-zero consent indicated the faculty member did not want to participate in all components associated with the program evaluation. A level-one consent indicated the faculty member would complete social validity forms (i.e., PIRS; Lane et al., 2002) twice per year and treatment integrity measures once per year (i.e., teacher self-report over time [TSR], SET). Finally, a level-

Table 3

Consents

School	Total number (n) of staff eligible for consent	Total number of consent level 0 (% of total eligible)	Total number of consent level 1 (% of total eligible)	Total number of consent level 2 (% of total eligible)	Total number of consent levels 1 and 2 (% of total eligible)	Total number of completed TSR (% of consented 1 and 2 completed)
A	52	3	8	41	49	38
B	14	0	3	11	14	11
C	34	0	4	30	34	30
D	31	0	7	24	31	16
E[a]	35	1	17	16	33	25
F	26	0	6	20	26	19
G[b]	49	0	7	39	46	23
H	29	0	12	17	29	21
Total	270	4 (1.48)	64 (23.70)	198 (73.33)	262 (97.04)	183 (69.85)

Note. [a]One consent form was not returned by an eligible consentee. [b]Three consent forms were not returned by eligible consentees.

Note. Consent level 0 = The faculty member will not participate in any evaluation components (e.g., PIRS, TSR, DO-T, SET); Consent level 1 = The faculty member will complete the PIRS once twice per year, TSR once per year, and SET (if selected) once per year; Consent level 2 = Consent Level 1 plus the teacher will be directly observed by a RA and complete the DO-T if randomly selected for observation.

two consent indicated the faculty member would complete everything at a level-one consent and, additionally, was willing to be directly observed by a RA and complete a treatment integrity form (i.e., teacher-completed observation rating [DO-T]) if he/she was randomly selected for observation. Completed consent forms were obtained from approximately 266 of the 270 eligible teachers and instructional staff. Of the 266 consented faculty members, 262 (i.e., consent level-one [*n* = 64] and consent level-two [*n* = 198]) agreed to complete the TSR, with 183 actually completing the TSR (see Table 3).

Treatment Integrity Measures

Development. The treatment integrity tool was developed based on measures used in a previous evaluation study (i.e., Lane, Kalberg, Bruhn, et al., 2008). As part of the previous evaluation study, primary plan treatment integrity of four of the schools from District A was assessed via direct observations and self-report using component checklists (Lane, Kalberg, Bruhn, et al.). These component checklists, which contained a 3-point Likert-type scale ranging from *not at all* (0), *part of the time* (1), to *all of the time* (2), included items reflecting Procedures for Teaching and Procedures for Reinforcing the primary plan at each individual school. These checklists, which were customized to the school's primary plan, served as the model for revising and developing the tool (i.e., TSR) used in the current study.

In the current study, the primary investigator and project coordinator held a meeting with participating RAs to design the teacher self-report (TSR) measure. The goal of this meeting was to create one tool that could be used universally across all schools implementing CI3T models. First, they examined the CI3T plans developed by each

school to identify components that were represented across all schools' plans. Then, they developed questions addressing these components (e.g., *Did I use behavior specific praise during student interactions?, Was my instruction linked to district/state standards?*). These questions were then placed into domains—Procedures for Teaching, Procedures for Reinforcing, and Procedures for Monitoring. Next, the group identified any redundant items and either combined or eliminated these items. They also separated complex items to form new, more specific items. The TSR consisted of 38 items: 16 Procedures for Teaching, 10 Procedures for Reinforcing, and 12 Procedures for Monitoring. Following the meeting, a lead research assistant (LRA) developed a rubric of operational definitions for each item and for each possible score on the Likert-scale (e.g., *no, not at all* [0], *yes, some of the time* [1], *yes, most of the time* [2], or *yes, all of the time* [3]). This scale was modified from the 3-point, Likert-type scale used in the previous evaluation study (Lane, Kalberg, Bruhn, et al., 2008) to a 4-point Likert-type scale, thus allowing for greater response specificity. The rubric also contained an additional column for notes of clarification.

Treatment integrity tools: Description of tools, scoring, and procedures. During the spring of the 2009-2010 school year, teachers with a consent level-one or -two for program evaluation (a) completed a treatment integrity checklist (TSR) to self-evaluate their implementation of the primary-level plan, and (b) may have been randomly selected to participate in SET interviews. Descriptions of the TSR and SET are included in the following paragraphs.

Teacher Self-Report (TSR). The TSR is a 38-item component checklist divided into three subscales--Procedures for Teaching, Reinforcing, and Monitoring school

expectations. Teachers rated themselves based on their implementation of the primary-level plan from the beginning of the current school year (i.e., Fall 2009) to the date of assessment (i.e., Spring 2010). The rating was based on a Likert-type scale ranging from *no, not at all* (0), *yes, some of the time* (1), *yes, most of the time* (2), or *yes, all of the time* (3). Percentage of implementation was calculated by dividing the total score by the total possible score and multiplying by 100. For example, if a teacher completed only 33 of 38 items, then the possible score was 99 (e.g., 33 x 3 = 99). Individual teacher percentages, mean teacher percentages by school, and a grand mean teacher percentage for all schools were calculated for the TSR.

The TSR forms were distributed into consent level-one and two teachers' school mailboxes by RAs and returned in a sealed envelope within each school. RAs collected the completed forms and returned them to Vanderbilt on the same day of collection. Scores on the TSR were entered into an EXCEL database by one RA, and reliability of entry was done by another RA on 30% of entry.

Schoolwide Evaluation Tool (SET). The SET, which has demonstrated internal consistency (Cronbach's alpha = .96) and test-retest reliability at 97.3% (Horner et al., 2004), was used to measure the SWPBS component of the CI3T. As previously described, the SET was designed to measure SWPBS implementation in seven different domains (i.e., Expectations Defined, Behavioral Expectations Taught, On-Going System for Rewarding Behavioral Expectations, System for Responding to Behavioral Violations, Monitoring and Decision-Making, Management, District-Level Support). RAs, which served as independent assessors, conducted SET interviews, observations, and material reviews one time during the same two-week period in the spring that the

TSR was distributed and returned. First, two RAs conducted an interview with an administrator. Then, a team of RAs interviewed at least 10 randomly selected teachers and 10% of the student population. This same team of RAs observed and reviewed various school materials (e.g., crisis prevention plan, office discipline referral form, school expectation posters). All SET data were returned to Vanderbilt on the same day as the assessment. When scoring the SET, the whole school (rather than an individual teacher) is the unit of analysis. The SET scoring guide provided in the SET manual was used to calculate scores. Calculating SET scores was done by one of two assigned RAs and 100% of data was made reliable by a third assigned RA.

Research assistant training procedures. RAs were provided explicit instruction on TSR and SET procedures by the project coordinator (PC) and a LRA during a 3 hr training session including modeling and guided practice. Specific items covered were: distribution and collection of TSR, directions for conducting the SET, and data entry and reliability procedures. Following the 3 hr training session, RAs independently practiced scoring the TSR and SET as well as entering scores into a mock database. Once the RA achieved 95% reliability on three of each form, he/she was deemed reliable and, thus, free to score and enter data independently. Further, RAs became reliable on SET interviews by shadowing a LRA who led (a) the administrator interview, (b) random teacher and student interviews, and (c) review of materials. Following interview completion, the LRA and RA discussed their separately recorded answers and resolved any discrepancies. The LRA determined when the RA was competent to complete future interviews independently.

Experimental Design and Statistical Analysis

The data analytic plan for examining the psychometric properties of the TSR was rooted in Classical Test Theory (CTT), which is a statistical theory based upon minimizing random measurement error (DeVellis, 2006; Nunnally & Bernstein, 1994). CTT is applied by measuring a tool's reliability (e.g., internal consistency) and validity (e.g., comparison with a credible, similar tool; DeVellis). More sophisticated models such as those grounded in Item-Response Theory (IRT) were not used for several reasons. First, application of CTT has been proven effective consistently throughout time (DeVellis, Nunnally & Bernstein). As Nunnally and Bernstein pointed out, a tool that is good by classical standards (i.e., CTT) will likely fit a suitable IRT model as well. Further, IRT requires a larger sample size (n = 200 to 500) than was available in this preliminary psychometric study. Factor analysis, which stems from CTT, was not conducted due to low sample size as factor analysis requires approximately 5 to 10 respondents per item (Nunnally & Bernstein).

To begin our application of CTT, we evaluated items from the TSR using descriptive statistics (i.e., mean, SD, skew, and kurtosis) and item-total correlations (Walker, Beck, Garber, & Lambert, 2009). Good items, as outlined by CTT, are those (a) free of floors or ceilings consequently limiting variance and exhibiting high skewness and kurtosis, and (b) having high item-total correlations yielding high alpha reliability (Walker et al.). Although the word *high* is used to describe item-total correlations, this is somewhat of a misnomer given typical test items usually correlate between .0 and .4 with any correlation above .2 generally considered a moderately discriminating item (Nunnally & Bernstein, 1994).

Examining descriptive statistics is just one step in what Benson (1998) termed the *structural stage* of the data analytic process used in validation studies. Specifically, the structural stage focuses on the internal relations among observed variables (e.g., internal consistency). The *external stage* focuses on relations among constructs (e.g., correlations with similar tools). Both stages of this validation study are described, in detail, in the following paragraphs.

Goal 1: Internal consistency. As part of the structural stage, to determine initial evidence of reliability of the TSR, the average correlation between items, or internal consistency, was analyzed using Cronbach's alpha (see Table 4). The alpha coefficient is an index of the extent to which instrument items measure the intended construct—in this case, treatment integrity of the CI3T primary plan. From a statistical perspective, an instrument demonstrates sufficient evidence of reliability if when the instrument is divided in half, the two halves are highly correlated (Cronbach, 1951). Essentially, Cronbach's alpha represents all possible correlations of the split-halves. An alpha of .70 is considered adequately reliable (Hatcher & Stepanski, 1994, Nunnally & Bernstein, 1994; Streiner, 2003), while .80 or higher is desirable (Nunnally & Bernstein, Streiner). However, when an instrument is used to make decisions about individuals (e.g., placement into intervention), .90 is the minimum and, .95 the ultimate goal (Nunnally & Bernstein). Alpha coefficients were computed for each subscale, or domain. Specifically for the TSR, we explored the internal consistency of the 16 items constituting Procedures for Teaching, 10 items constituting Procedures for Reinforcing, and 12 items constituting Procedures for Monitoring.

Consistent with CTT associated methods for test construction, items were considered for removal based upon empirical and theoretical evidence. We begin with defining the empirical criteria. Specifically, to make decisions at the item level we considered three empirical criteria: (a) Did the item demonstrate a floor or ceiling effect (Walker et al., 2009)?, (b) Was the item-total correlation less than or equal to .20 (Nunnally & Bernstein)?, and (c) Does removing an individual item from a subscale improve alpha values (Hatcher & Stepanski)?. First, a floor or ceiling effect was detected by examining the mean of each item. A mean close to 0 or 3 indicated a floor or ceiling, respectively, and offered potential evidence the item contributed little valuable information (Walker et al). Second, the item-total correlation was evaluated to determine if the item was discriminating, or in other words, the item could sharply discriminate between those who scored low or high on the total subscale (DeVellis). Finally, for example, if removing an item raised the overall alpha from .72 to .84, it was considered for removal, as this presented evidence the item was not measuring the same construct as the other included items (Hatcher & Stepanski).

Because alpha values are based on interitem correlations, we examined the, intercorrelations between within-subscale items as well as intercorrelations between subscales using Pearson product-moment correlation coefficients. To meet the assumptions required to accurately interpret alpha values, interitem (or intersubscale) correlations should not differ substantially, thus demonstrating independence (Cronbach & Shavelson, 2004; Vincent et al., 2010). Thus, we examined the range, level, and significance of interitem and intersubscale correlations. It is important to note, however, the assumption of independence is rarely met. Moreover, assessing the degree and effect

of non-independence is not only cumbersome, but nearly impossible (Cronbach & Shavelson).

After examining empirical criteria and prior to item removal, we considered our knowledge of core components of CI3T models. For example, a key component to implementing the CI3T plan via reinforcing procedures is delivering tickets to students for meeting schoolwide academic, behavioral, and social expectations (Lane et al., in press). Because this component is a cornerstone of implementation, it was not considered for removal. It is important to note, however, item removal was based upon both empirical and theoretical knowledge. Therefore, it was possible removing an item improved the alpha coefficient (i.e., empirical knowledge), but it remained in the instrument because our understanding of primary plan constructs indicated it was an essential component (i.e., theoretical knowledge). Ultimately, decisions about item removal were based on human judgment using multiple criteria including (a) those first derived empirically (as previously described) and then (b) those based upon our theoretical understanding of components essential to CI3T implementation (Nunnally & Bernstein). By considering both empirical and theoretical evidence, we were able to make informed, balanced decisions about item removal.

Goal 2: Correlations between the TSR and SET. As outlined by CTT, one goal of validation studies is to demonstrate assessment scores are consistent with our theoretical understanding of how the construct of interest truly occurs in the real world (DeVellis, 2006). Benson called this the external stage of analysis (1998). Often, this goal is achieved through comparing the relationship of a new assessment (e.g., TSR) with scores obtained on credible measures (e.g., SET; DeVellis). Therefore, we examined

evidence of convergent and divergent validity by examining the correlations between the TSR and SET (Benson; DeVellis). Because the TSR and the SET (a) provide an index of how the primary plan has been implemented over time, and (b) measure similar behavioral components of the primary plan; the relation between the two measures was analyzed. Given the SET uses the school as the unit of analysis and the TSR uses the teacher as the unit of analysis, the TSR had to be aggregated at the school level. Specifically, school means on the TSR were calculated by averaging the teacher scores thus creating eight school means for each subscale and total score. Next, a grand mean for each subscale and total was calculated by averaging the eight school means on each subscale and total. Then, like other studies of evidence based on relations to similar assessments (e.g., Vincent et al., 2010), our first set of data analyses was done at the domain (i.e., subscale) level. To be precise, the TSR has three domains or subscales—Procedures for Teaching, Procedures for Reinforcing, and Procedures for Monitoring—and a total score which were analyzed according to their Pearson product-moment correlations with the seven SET subscales (i.e., Expectations Defined, Behavioral Expectations Taught, On-Going System for Rewarding Behavioral Expectations, System for Responding to Behavioral Violations, Monitoring and Decision-Making, Management, District-Level Support) and total score (see Table 8). Preliminary examination of the relation between the SET and TSR subscales and total scores indicated a lack of significant positive correlations despite the fact that both the SET and TSR measure primary plan behavioral components (see Table 8).

In an effort to better understand our findings, we conducted a second analysis similar to Vincent, Spaulding, and Tobin (2010) who rearranged items on subscales to

perform a more logical comparison of two instruments (i.e., the SET and TIC). Namely, although the TSR and SET share similar behavioral items, the arrangement of subscales is not identical. After reading through each item constituting subscales on both measures, we determined some SET subscales could be combined to form a new subscale that was more comparable to TSR subscales. Thus, we aggregated SET subscales to create subscales more closely aligned with TSR subscales (Vincent et al., 2010) to form the SET-NEW (see Table 9). For example, SET subscales Expectations Defined and Behavioral Expectations Taught were combined to form the SET-T as the items constituting these subscales were similar in construct and wording to the behavioral items constituting the TSR Procedures for Teaching subscale (e.g., SET: "*Is there a documented system for teaching behavioral expectations to students on an annual basis?*, and TSR*: "Did my students receive instruction about our schoolwide expectations for each setting?").* On-Going System for Rewarding Behavioral Expectations and System for Responding to Behavioral Violations subscales were combined to form the SET-R, as they were comparable to the TSR-Procedures for Reinforcing subscale (e.g., SET: *"Is there a documented system for rewarding student behavior?"*, and TSR: *"Did I give tickets to students demonstrating schoowide expectations?"*). Finally, Monitoring and Decision-Making and Management subscales formed the SET-M, which was similar to the TSR Procedures for Monitoring subscale (e.g., SET: *"Do 90% of team members asked report that discipline data are used for making decisions in designing, implementing, and revising schoolwide effective behavior support efforts?"*, and TSR: *"Did I use behavioral data to inform my instruction of at risk students?"*). Like the study by Vincent and colleagues, the District-Level Support subscale remained the same, and

thus was not redistributed because no items in any TSR subscale addressed district support. Similar to the first analysis of the SET and TSR, means and standard deviations on the SET-T, SET-R, and SET-M were calculated for the SET-NEW. Then, the SET-NEW subscales (i.e., SET-T, SET-R, SET-M), District-Level Support subscale (i.e., SET-DLS), and SET total score were analyzed for their correlation with the TSR subscales and total score (see Table 9). When interpreting correlations in the first and second set of analyses, coefficients of .20 or less were considered weak, .50 were moderate, and .80 and higher were strong (Hatcher & Stepanski, 1994).

CHAPTER III

RESULTS

Reliability

Item-level analyses. First, we examined item-level means on each subscale. Means ranged from 1.69 (Item 2: *Setting expectations posted*) to 2.84 (Item 16: *Clear routines*) on Procedures for Teaching. Procedures for Reinforcing means ranged from 1.55 (Item 10: *Used tickets to facilitate routines*) to 2.76 (Item 7: *Refrained from taking ticket away*). Finally, Procedures for Reinforcing means were between 1.71 (Item 12: *Made referrals for students shy/withdrawn*) and 2.82 (Item 4: *Administered academic progress monitoring assessments*).

Further analysis revealed all items on all subscales, with the exception of Procedures for Teaching Item 12 (*Positive tone during student interactions*) which was perfectly normally distributed (skew = 0.00), were slightly negatively skewed ranging from -3.49 (Teaching Item 1: *3-5 schoolwide expectations posted, visible*) to -.16 (Reinforcing Item 10: *Used tickets to facilitate routines*). Given item-level means were all fairly high, with several demonstrating ceiling effects (mean > 2.75; see Table 4) and none near the floor (mean < .50), the negative skew was expected. For example, the mean for Procedures for Teaching Item 1 which asks "*Did I have our 3 to 5 schoolwide expectations posted and visible in my classroom?*" was 2.82 and had a -3.49 skew indicating this item was often endorsed as the majority of participants scored near the top of the scale (i.e., 3). Examination of kurtosis indicated variability with values as low as -

1.82 (i.e., platykurtic, or flat distribution) and as high as 13.34 (i.e., leptokurtic, or peaked distribution).

Table 4
TSR: Descriptive Statistics and Cronbach's Coefficient Alphas

Subscale	*M* > 2.75	*SD*	Skew	Kurtosis > 9.00	Standardized Variables: *r* With Total ≤ .20	Standardized Variables: Alpha
Procedures for Teaching						.83
1 3-5 schoolwide expectations posted, visible	2.82	0.65	-3.49	11.06	.20	.83
2 Setting expectations posted	1.69	1.36	-0.18	-1.82	.33	.82
3 Instruction on setting expectations	2.35	0.86	-1.13	0.32	.46	.81
4 Instruction on social skills	2.18	0.90	-0.91	-0.11	.49	.81
5 Modeled behavioral expectations	2.75	0.48	-2.15	5.37	.49	.81
6 Instruction linked to district/state standards	2.83	0.45	-2.83	9.81	.31	.82
7 Differentiated academic instruction	2.67	0.54	-1.48	1.29	.58	.81
8 Made social or behavioral modifications	2.68	0.50	-1.77	3.94	.49	.81
9 Engaged students beginning to end of class	2.45	0.53	-0.24	-1.22	.50	.81
10 Conducted daily starting activities	2.54	0.70	-1.75	3.25	.41	.82
11 Conducted daily closing activities	2.35	0.78	-1.14	0.84	.50	.81
12 Positive tone during student interactions	2.38	0.53	0.00	-0.93	.44	.82
13 Procedures to foster safe environment	2.77	0.49	-2.09	3.56	.27	.83
14 Support for students who missed instruction	2.68	0.53	-1.75	3.71	.56	.81
15 Checked for understanding on directions	2.77	0.42	-1.48	0.77	.56	.81
16 Clear routines for class procedures	2.84	0.39	-2.08	3.23	.34	.82
Procedures for Reinforcing						.76
1 Delivered school's reactive plan consequences	2.63	0.55	-1.39	1.83	.50	.73

Table 4, Continued

Subscale	*M* > 2.75	*SD*	Skew	Kurtosis > 9.00	Standardized Variables: *r* With Total ≤ .20	Standardized Variables: Alpha
3 Gave behavior specific praise	2.62	0.51	-0.72	-0.90	.56	.72
4 Behavior specific praise when giving tickets	2.73	0.46	-1.14	-0.25	.47	.73
5 Allowed ticket exchange for rewards	2.63	0.79	-2.33	4.75	.28	.76
6 Allowed participation in schoolwide drawings	2.63	0.80	-2.29	4.49	.34	.75
7 Refrained from taking tickets away	2.76	0.70	-3.40	11.10	.24	.76
8 Received positive feedback from colleagues or administrators	2.05	0.90	-0.76	-0.15	.52	.72
9 Perception of school's plan favorable amongst colleagues and administrators	2.18	0.67	-0.44	-0.15	.40	.74
10 Used tickets to facilitate routines	1.55	1.20	-0.16	-1.49	.35	.75
Procedures for Monitoring						.85
1 Filled out discipline referrals	2.05	1.00	-0.81	-0.39	.33	.85
2 Completed behavior screeners	2.64	0.80	-2.49	5.46	.36	.84
3 Completed attendance procedures	2.75	0.74	-3.29	9.80	.43	.84
4 Administered academic progress monitoring assessments	2.82	0.60	-3.65	13.34	.31	.85
5 School shared school-wide behavior data	2.56	0.85	-1.60	1.44	.41	.84
6 School shared school-wide academic data	2.64	0.72	-2.06	3.81	.54	.83
7 Used behavior data to inform instruction	2.28	0.88	-1.10	0.40	.65	.82
8 Used academic data to inform instruction	2.58	0.74	-2.12	4.23	.60	.83
9 Used behavior and academic data together	2.35	0.77	-1.24	1.00	.68	.82
10 Made referrals for students struggling academically	2.34	0.98	-1.29	0.40	.69	.82
11 Made referrals for students acting out	2.24	1.02	-1.13	0.03	.62	.83
12 Made referrals for students shy/withdrawn	1.71	1.20	-0.25	-1.51	.53	.83

In Table 4, we underlined items demonstrating (a) means near the ceiling (i.e., greater than 2.75), (b) excessive kurtosis (i.e., greater than 9.00), and (c) with low item-total correlations (i.e., $r \leq .20$). These items were flagged for possible removal under CTT, which asserts good items are free of (a) floors or ceilings and (b) extreme skewness and kurtosis, as well as demonstrating item-total correlations above at least .20 (Nunnally & Bernstein, 1994; Walker et al., 2009). Due to the homogeneity of skew values, that is, all items were slightly negatively skewed; no criterion for skew was included.

Additionally, interitem correlations were calculated within each subscale and ranged from -.06 (Item 1: *3 to 5 schoolwide expectations posted*, Item 9*: Engaged students beginning to end of class*) to .72 (Item 3: *Instruction on setting expectations*, Item 4: *Instruction on social skills*) on Teaching items, .02 (Item 5: *Allowed ticket exchange*, Item 9: *Perception of school's plan*) to .61 (Item 3: *Gave behavior specific praise*, Item 4: *Gave behavior specific praise when giving tickets*) on Reinforcing items, and -.05 (Item 1: *Filled out discipline referrals*, Item 4: *Administered academic progress monitoring assessments*) to .70 (Item 8: *Used academic data to inform instruction*, Item 9: *Used behavior and academic data together*) on Monitoring items (see Tables 5, 6, 7). All interitem correlations were not significant. Specifically, there were several interitem correlations on each subscale (e.g., Teaching Item 1 with Teaching Items 6, 7, 8, 9, 10, 11, 12, 14, 14, 16) that were near zero but were not statistically significant, and thus, we could not be sure if they occurred by chance. Many significant correlations on items within subscales were significant, however. For example, Teaching Item 6 (*Instruction linked to district/state standards*) and Teaching Item 7 (*Differentiated academic instruction*) were significantly correlated ($r = .54$, $p < .0001$); as were (a) Reinforcing

Item 2 (*Gave tickets to students meeting expectations*) and Reinforcing Item 6 (*Allowed participation in schoolwide drawings; r* = .27, *p* < .001), and (b) Monitoring Item 4 (*Administered academic progress monitoring assessments*) and Monitoring Item 8 (*Used academic data to inform instruction*; *r* = .30, *p* < .001). Significant low to moderate

Table 5

Interitem Correlations: Procedures for Teaching

Item	1	2	3	4	5	6	7	8
1	1.00							
2	.29***	1.00						
3	.36****	.35****	1.00					
4	.32***	.18*	.72****	1.00				
5	.18*	.28***	.40****	.39****	1.00			
6	-.03	.08	.08	.13	.17*	1.00		
7	.06	.10	.18	.23**	.33****	.54****	1.00	
8	.09	.05	.11	.23**	.19*	.30***	.61****	1.00
9	-.06	.15	.10	.15	.21*	.21*	.41****	.44****
10	.02	.14	.30***	.31***	.25**	.11	.16	.14
11	.01	.23**	.26**	.31***	.29***	.19*	.29***	.10
12	.01	.30***	.06	.18*	.28***	.09	.25**	.27**
13	.00	.11	.14	.10	.16	.18*	.21*	.17*
14	.06	.19*	.21*	.31***	.29***	.16	.48****	.42****
15	.24**	.11	.16	.19*	.26**	.21*	.46****	.51****
16	-.03	.13	.19*	.10	.18*	.05	.20*	.23**

* p<.05, ** p<.01, *** p<.001, **** p<.0001

Table 5, Continued

Item	9	10	11	12	13	14	15	16
1								
2								
3								
4								
5								
6								
7								
8								
9	1.00							
10	.27**	1.00						
11	.29***	.72****	1.00					
12	.49****	.14	.26**	1.00				
13	.15	.02	.14	.26**	1.00			
14	.39****	.31***	.38****	.30***	.19*	1.00		
15	.44****	.15	.25**	.36****	.23**	.42****	1.00	
16	.32***	.18*	.21*	.19*	.15	.29***	.35****	1.00

* p<.05, ** p<.01, *** p<.001, **** p<.0001

Table 6

Interitem Correlations: Procedures for Reinforcing

Item	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	.38****	1.00								
3	.42****	.56****	1.00							
4	.33****	.44****	.61****	1.00						
5	.09	.18*	.17**	.10	1.00					
6	.14	.27***	.22**	.22**	.34****	1.00				
7	.32****	.03	.16*	.06	.17*	.08	1.00			
8	.33****	.27***	.23*	.26**	.27***	.22**	.16*	1.00		
9	.29***	.39***	.24**	.23**	.02	.02	.22**	.52****	1.00	
10	.25**	.27***	.17*	.14	.15	.25**	.07	.39****	.13	1.00

* p<.05, ** p<.01, *** p<.001, **** p<.0001

Table 7

Interitem Correlations: Procedures for Monitoring

Item	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00											
2	.23**	1.00										
3	.08	.56****	1.00									
4	-.05	.21*	.42****	1.00								
5	.22**	.12	.06	.11	1.00							
6	.37****	.10	.15	.11	.61****	1.00						
7	.24**	.18*	.31***	.16	.33****	.48****	1.00					
8	.21**	.28***	.27***	.30***	.25**	.38****	.47****	1.00				
9	.26**	.20*	.17*	.28***	.38****	.43****	.69****	.70****	1.00			
10	.13	.26**	.40****	.37****	.28***	.39****	.49****	.52****	.54****	1.00		
11	.30***	.17*	.27**	.17*	.30***	.34****	.45****	.35****	.44****	.67****	1.00	
12	.31***	.16**	.23**	.06	.18*	.24**	.50****	.26**	.44****	.51****	.65****	1.00

* p<.05, ** p<.01, *** p<.001, **** p<.0001

interitem correlations such as the examples provided indicated items were related but uniquely contributed to measuring the overall constructs of teaching, reinforcing, and monitoring the CI3T primary plan; whereas significantly high correlations would have indicated items were not uniquely contributing and, thus, redundant (Streiner, 2003).

In examining subscale correlations (see Table 8), we found Teaching and Reinforcing subscales on the TSR were significantly and moderately correlated ($r = .74$, $p < .05$) as were Monitoring and Reinforcing ($r = .71$, $p < .05$). While Monitoring and Teaching demonstrated a moderate correlation of .61, it was not significant. All TSR subscales were significantly, highly correlated with the TSR total score. The Teaching and Total Score correlation was .91 ($p < .01$), Reinforcing and Total was .91 ($p < .01$), and Monitoring and Total was .80 ($p < .05$), thus indicating all subscales contributed to the TSR total score.

Internal consistency. Each subscale on the TSR (i.e., Procedures for Teaching, Reinforcing, Monitoring) yielded Cronbach's alpha above .70, with two of three subscales exceeding the desired .80 (Hatcher & Stepanski; Nunnally & Bernstein; Streiner). However, none met .90, which is the minimum value recommended for instruments used to make differential decisions (Nunnally & Bernstein). According to our analysis, removing any item on each subscale would not have improved overall subscale alpha values. For example, removing any Teaching item would have yielded the same overall alpha of .83 (e.g., Item 1: *3 to 5 schoolwide expectations posted*) or decreased it to as low as .81 (e.g., Item 14: *Support for students missing instruction*). Similarly, removing any Monitoring item would have kept alpha at .76 (e.g., Item 5: *Allowed ticket exchange for rewards*) or decreased it to as low as .72 (e.g., Item 3: *Gave*

specific praise). We coupled this information with item-level analysis, which yielded only one item (Teaching Item 1: *3 to 5 schoolwide expectations posted*) meeting multiple empirical criteria (i.e., ceiling effect, excessive skew and kurtosis, and low item-total correlation) for removal consideration. However, given our knowledge of CI3T plans, we deemed Teaching Item 1 essential to primary plan implementation from a theoretical perspective. Namely, every classroom in the building should have the schoolwide expectations posted and visible because (a) this is one way teachers and instructional staff communicate expectations to students and (b) they provide a visual reminder to students. Therefore, in considering both empirical and theoretical evidence, we elected to leave all items in the TSR. Thus, the alpha value for the Procedures for Teaching subscale equaled .83, the Procedures for Reinforcing subscale value equaled .76, and the Procedures for Monitoring subscale value equaled .85 (see Table 4).

Validity

TSR and SET subscale and total correlations. In our first set of validity analyses, we examined correlations between the TSR and SET for significance and magnitude (see Table 8). Correlations ranged from -.69 (SET-SFR and TSR-R) to .17 (SET-MDM and TSR-R), with the vast majority being low to moderate, negative correlations (e.g., TSR-T and SET-OGS, r = -.49). Negative correlations indicated as SET scores increased, TSR scores decreased and vice versa. Two correlations, (a) TSR-M and SET-BET (r = .04) and (b) TSR-T and SET-DLS (r = -.03) were near zero indicating no relationship. Only two correlations were positive, albeit low magnitude (TSR-R and SET-MDR: r = .17; TSR-M and SET-MDR: r = .13) indicating subscales were slightly, but not significantly, related.

Table 8

SET and TSR Correlations

Measure	*M* (*SD*)	1	2	3	4	5	6	7	8	9	10	11	12
1. SET-ED	96.86 (8.83)	1.00											
2. SET-BET	96.25 (7.44)	-.20	1.00										
3. SET-OGS	97.92 (5.89)	1.00**	-.20	1.00									
4. SET-SFR	82.81 (13.26)	.24	.70*	.24	1.00								
5. SET-MDM	84.38 (11.08)	.34	.49	.34	.19	1.00							
6. SET-MA	82.12 (7.28)	.78*	.12	.78*	.43	.21	1.00						
7. SET-DLS	93.75 (17.68)	-.14	-.20	-.14	-.14	-.11	-.26	1.00					
8. SET-TOT	89.47 (7.91)	.86**	.21	.86**	.55	.54	.76*	.10	1.00				
9. TSR-T	85.20 (3.06)	-.49	-.29	-.49	-.55	-.43	-.33	-.03	-.65	1.00			
10. TSR-R	81.04 (5.54)	-.20	-.25	-.20	-.69	.17	-.21	-.23	-.41	.74*	1.00		
11. TSR-M	78.91 (3.75)	-.37	.04	-.37	-.56	.13	-.37	-.26	-.49	.61	.71*	1.00	
12. TSR-TOT	82.18 (3.51)	-.37	-.29	-.37	-.67	-.14	-.37	-.21	-.60	.91**	.91**	.80*	1.00

Note. The SET subscales and total are presented in items 1-8 (i.e., SET-ED = *Expectations Defined*, SET-BET = *Behavioral Expectations Taught*, SET-OGS = *On-Going System for Rewarding Behavioral Expectations*, SET-SFR = *System for Responding to Behavioral Violations*, SET-MDM = *Monitoring and Decision-Making*, SET-MA = *Management*, SET-DLS = *District-Level Support*, SET-TOT = *Total Score*). The TSR subscales and total are presented in items 9-12 (i.e., TSR-T = *Procedures for Teaching*, TSR-R = *Procedures for Reinforcing*, TSR-M = *Procedures for Monitoring*, TSR-TOT = *Total Score*).

* p<.05, ** p<.01, *** p<.001, **** p<.0001

Overall, results indicated no SET subscales were significantly correlated with any TSR subscales. Although the TSR total and SET total scores were moderately correlated, these also were insignificant. Given (a) the lack of significant correlations between the TSR and SET subscales and total scores and (b) our knowledge that both the TSR and SET measure primary plan behavioral components; we decided to conduct another set of analyses following the work of Vincent, Spaulding, and Tobin (2010).

TSR and SET-NEW subscale and total correlations. After redistributing domains on the SET to form fewer subscales containing more items (i.e., SET-NEW; see Table 9) that were somewhat parallel to TSR subscales, we examined the SET-NEW and TSR subscale and total correlations. Again, nearly all correlations were moderate and negative ranging from -.71 ($p < .10$; SET-R and TSR-TOT) to -.09 (SET-M and TSR-M), with two correlations near zero (TSR-T and SET-DLS: $r = -.03$; TSR-R and SET-M: $r = .03$). However, unlike the first comparison between the TSR and SET, these analyses of the TSR and SET-NEW yielded significant correlations between some subscales and the total. For example, the TSR-T was moderately correlated with the SET-T ($r = -.62$, $p < .10$), SET-R ($r = -.64$, $p < .10$), and SET total ($r = -.65$, $p < .10$). Thus, as TSR-T scores increased, SET-T, SET-R, and SET total scores decreased and vice versa. Parallel findings were found with the SET-R which was significantly and moderately correlated with the TSR-R ($r = -.66$, $p < .10$) and the TSR total ($r = -.71$, $p < .10$).

Table 9

SET-NEW and TSR Correlations

Measure	*M* (*SD*)	1	2	3	4	5	6	7	8	9
1. SET-T	96.56 (5.16)	1.00								
2. SET-R	90.36 (7.87)	.86**	1.00							
3. SET-M	83.59 (7.23)	.87**	.55	1.00						
4. SET-DLS	93.75 (17.68)	-.27	-.17	-.22	1.00					
5. SET-TOT	89.47 (7.91)	.88**	.78*	.80*	.10	1.00				
6. TSR-T	85.20 (3.06)	-.62+	-.64+	-.50	-.03	-.65+	1.00			
7. TSR-R	81.04 (5.54)	-.35	-.66+	.03	-.23	-.41	.74*	1.00		
8. TSR-M	78.91 (3.75)	-.29	-.61	-.09	-.26	-.49	.61	.71*	1.00	
9. TSR-TOT	82.18 (3.51)	-.52	-.71+	-.29	-.21	-.60	.91**	.91**	.80*	1.00

Note. The SET-NEW subscales (SET-T = *Expectations Defined* and *Behavioral Expectations Taught*; SET-R = *On-Going System for Rewarding Behavioral Expectations* and *System for Responding to Behavioral Violations*; SET-M = *Monitoring and Decision-Making* and *Management)*, SET-DLS = *District-Level Support*, and SET-TOT = *SET total score* are presented in items 1-5. The TSR subscales and total are presented in items 6-9 (i.e., TSR-T = *Procedures for Teaching*, TSR-R = *Procedures for Reinforcing*, TSR-M = *Procedures for Monitoring*, TSR-TOT = *Total Score*).

+p <.10, * p<.05, ** p<.01, *** p<.001, **** p<.0001

CHAPTER IV

DISCUSSION

Across the country, schools are using three-tiered prevention models such as RtI, SWPBS, and CI3T to support the academic, behavioral, and/or social needs of students. These models are grounded in using data to make decisions about student responsiveness to varying levels of intervention. Understanding the precision with which an intervention is being implemented is central to this process (Schulte et al., 2009). While assessing all procedural variables during baseline (or control) and treatment conditions (i.e., procedural fidelity) would provide the most thorough understanding of student responsiveness to CI3T implementation (Wolery & Ledford, 2011), we focused our study on the procedural variables associated with the treatment condition only (i.e., treatment integrity). This initial focus on treatment integrity, rather than procedural fidelity, aligns with experts' current development and evaluation of multi-tiered prevention models. Namely, our field is in the beginning stages of (a) developing psychometrically sound treatment integrity assessments, and (b) learning how to use these data to make sound instructional and administrative decisions. To begin the evaluative process, at a minimum, we must collect treatment integrity data because only when the degree of implementation is known can accurate decisions about student responsiveness be made (Gersten et al., 2005; Gresham et al., 1993; Horner et al., 2005). Without these data, decisions may render unnecessary or incorrect student placement. For example, a student may exhibit problem behavior in a classroom and be referred for a tier two intervention.

Prior to placing the student in intervention, schools need to ensure the student was effectively exposed to the primary prevention plan (Bruhn et al., 2011). If not, it is possible a more accurately implemented primary plan would be sufficient in reducing the problem behavior (Bruhn et al.). Because schools have finite resources allocated to interventions, it is essential only students who truly need intervention receive it. To make this determination, schools must measure primary plan treatment integrity.

Several tools exist for measuring primary plan treatment integrity of SWPBS—the behavioral framework used in CI3T models. Most of these tools involve self- or team-assessment. And, the self- or team-assessments reflect impressions of whole-school implementation rather than implementation in individual classrooms. Unfortunately, these tools are limited by their inability to capture classroom-level implementation by individual teachers and other instructional staff. As schools move forward in three-tiered prevention models, the quality of their decisions on student responsiveness hinges on the quality with which the prevention plans are carried out by each teacher in the building. Therefore, evaluation of implementation must occur in individual classrooms using measures allowing reliable and valid inferences to be drawn and quality decisions to be made.

To this end, we developed a classroom-level tool to measure treatment integrity of the primary prevention plan implemented within a CI3T model encompassing academic, behavioral, and social components. Before this tool can be used for making data-based decisions such as determining student responsiveness to the primary plan and allocating professional development resources for teachers who may need more implementation

support, we needed to ensure the tool met strong psychometric standards by consistently producing accurate results.

Reliability

The first goal of the study focused on establishing reliability of the TSR using methods derived from CTT. To begin the process of the structural stage (Benson, 1998), we examined item-level descriptive statistics such as mean, standard deviation, skew, kurtosis, and item-total correlations. These statistics provided the empirical evidence used to make decisions about item retention. Although some items demonstrated ceiling effects (mean > 2.75), low item-total correlations ($r < .20$), and excessive kurtosis (> 9.00); only one item (i.e., Teaching Item 1: *3 to 5 schoolwide expectations posted*) demonstrated all of these characteristics. Because it was essential to consider theoretical, in addition to empirical, evidence; we elected to keep Teaching Item 1 as we believed it critical to CI3T implementation. Further, a congruent item exists on the SET (i.e., *"Are the agreed upon rules & expectations publicly posted in 8 of 10 locations?"*), indicating the item is, in fact, a core behavioral component to the primary plan.

In addition to item-level analyses, we examined subscale correlations within the TSR. Interestingly, Teaching and Reinforcing subscales were significantly and moderately correlated ($r = .74$, $p < .05$) as were Monitoring and Reinforcing ($r = .71$, $p < .05$). The Teaching and Reinforcing subscale correlation was expected given school teams were taught during CI3T training to reinforce students for meeting academic, behavioral, and social expectations and content that had been taught in the classroom. The correlation between Monitoring and Reinforcing, however, was unforeseen given the

Monitoring subscale items mostly reflected how academic and behavioral data were used, while the Reinforcing subscale addressed how positive reinforcement and reactive discipline were delivered to students. Rather, we expected significant and moderate correlations between Monitoring and Teaching because the Teaching subscale addressed differentiated instruction and student modifications and the Monitoring subscale contained items associated with using data to inform instructional decisions (e.g., curricular modifications). Overall, the moderate subscale correlations indicated the subscales were related but not redundant, thus demonstrating adequate evidence these subscales contributed to the TSR total score (Horner et al., 2004, Streiner, 2003). This was further demonstrated by each subscale's high and significant correlation with the TSR Total Score.

Next, we evaluated reliability of the TSR containing all original items (as all were retained) using an internal consistency estimate, which is the degree to which individual items correlate with each other or the total as measured by coefficient alpha (Hatcher & Stepanski, 1994). Further analysis confirmed our decision to retain all items, as removing an item would not have improved overall alpha values on any subscale (Hatcher & Stepanski). Thus, we concluded all items were measuring the intended constructs (i.e., teaching, monitoring, and reinforcing the CI3T primary plan; Hatcher & Stepanski). Results indicated acceptable ($\alpha = .76$, Procedures for Reinforcing) to good ($\alpha = .83$, Procedures for Teaching; $\alpha = .85$, Procedures for Monitoring) internal consistency for scale items (Hatcher & Stepanski; Nunnally & Bernstein, 1994, Streiner, 2003). We speculated that if the data had been more variable, then higher alpha values may have been obtained (Nunnally & Bernstein). Namely, given the negative skew due to many

high scores, our data may have lacked the variability necessary to obtain higher alpha values. It is possible the lack of variability resulted from what Nunnally and Bernstein termed *mastery learning*, which manifests when instruction has the desired effects. In this case, because schools were trained by field experts in effective CI3T implementation practices, it is possible participants were, as they reported, implementing at high levels. However, we must consider the TSR utilizes self-report. As researchers have shown, self-report data tend to be a bit inflated in comparison to data obtained by outside observers making it difficult to determine how the true score actually differed from the observed (i.e., self-reported) score (Lane, Kalberg, Bruhn, et al., 2008). Therefore, an important next step in assessing treatment integrity may include developing tools that can be used by external assessors to conduct direct observations of classroom implementation.

On the other hand, alpha coefficients tend to be higher when items have maximally similar distributions (Nunnally & Bernstein). In this case, all but one item was slightly negatively skewed indicating they had very similar distributions. Thus, it is conceivable alpha values obtained in this initial analysis reached their potential magnitude. Although alpha coefficients were not as high in comparison to other treatment integrity tools measuring school-level implementation (e.g., SET: $\alpha = .96$; Horner et al., 2004; BoQ: $\alpha = .96$; Cohen et al., 2007), findings suggested the TSR demonstrated initial evidence of an adequately reliable classroom-level tool for measuring primary plan treatment integrity in a CI3T model. However, because assessment tools used to make decisions about individuals should demonstrate alpha values above .90 (Nunnally & Bernstein); the TSR likely needs further modifications with subsequent analyses to render it acceptable for making decisions about providing (a)

support to teachers implementing at low levels, and (b) intervention to students non-responsive to the primary plan. Thus, based on these initial internal consistency estimates, schools should not rely on the TSR alone to determine if the primary plan is being implemented accurately by teachers and instructional staff. Instead, the TSR should be used either in conjunction with other tools measuring different perspectives (e.g., SET) or as a formative assessment.

Validity

The second goal of this study was to examine the relationship between the TSR and SET by computing correlations between each subscale and total score on the TSR with each subscale and total score on the SET. Although we hypothesized TSR subscales containing salient behavioral items would be correlated with SET subscales, none of the TSR and SET subscales and total scores were significantly, positively correlated. For example, SET subscales Expectations Defined and Behavioral Expectations Taught contained questions regarding the teaching of schoolwide behavioral expectations, and the TSR subscale Procedures for Teaching contained several similar items. Yet, these subscales were not significantly, positively correlated. Rather, they were negatively correlated, although not significantly. This was surprising given the TSR and SET assess overlapping information, and thus, we had anticipated the subscales with salient behavioral items to converge. However, because the TSR was designed to measure not only the schoolwide behavior plan as the SET does, but also academic and social components, these additional components likely contributed to the insignificance of the correlations providing some initial evidence the SET and TSR are tapping different

constructs. Specifically, they yielded independent, rather than overlapping, information (i.e., divergent validity). Further, it is also possible the lack of significant correlations between SET and TSR subscales may be attributed to different sources of data used in the assessment process. Namely, the SET is conducted by external assessors at the school level through a series of interviews and review of materials, while the TSR is completed by teachers and instructional staff at the classroom level, and thus, the SET may not capture teacher implementation in individual classrooms. Most likely, however, these unexpected findings may be attributed to the small sample (n = 8) and range restriction of SET scores (i.e., all schools scored near the top of the scale creating little variability) which limited the statistical power to detect significant correlations.

In light of these findings, we conducted additional analyses to determine if combining subscales to form more parallel measures would produce convergent, rather than divergent, evidence of validity. Similar to the TSR and SET comparison, all TSR and SET-NEW correlations were either negative or near zero. These analyses, however, yielded a few significant correlations such as the TSR-T with the SET-T (r = -.62, p < .10), SET-R (r = -.64, p < .10), and SET total (r = -.65, p < -.65). Again, these findings were not expected particularly because it was counterintuitive to think subscales containing overlapping information would be negatively correlated. Like our first analyses of TSR and SET correlations, we attributed these findings to (a) the small sample size, (b) the lack of variable SET scores due to all schools scoring near the top of the scale, (c) the additional academic and social component items included in the TSR, and (d) the different sources of data (i.e., external assessment versus self-report) utilizing different units of analysis (i.e., school versus teacher). Because both sets of analyses

indicated the SET and TSR were more different than similar, it is possible both instruments generate unique information that would be lost if only one of the two instruments were used to assess treatment integrity of the primary plan. However, definitive conclusions about the relationship between the TSR and SET or SET-NEW must be tempered by the substantial limitations described below.

Limitations and Future Directions

This first validation study of the TSR indicated good internal consistency, as well as divergent validity with the SET. Yet, there are several important limitations associated with the reliability and validity findings. First, in terms of validity, our examination of the relationship between the SET and the TSR was severely limited by sample size (n = 8). Although 183 teachers and instructional staff completed the TSR, because the school is the unit of analysis for the SET, the TSR had to be aggregated at the school level for validity comparisons. This small sample size contributed to the lack of variability in the SET data, which was compounded by the fact that SET scores generally lack variability (Vincent et al., 2010). As Vincent and colleagues pointed out, the SET effectively assesses initial implementation, but may not reflect gradual changes in implementation over time. Therefore, SET scores lack variability regardless of the number of years a school has implemented SWPBS. Examination of our raw data yielded similar findings with means on all subscales near the ceiling (i.e., 100%, see Tables 8 and 9). In light of the small sample size and absence of variability in the SET data, the statistical power to detect meaningful correlations between the SET and TSR was low. Undoubtedly, future research should include a much larger sample of schools

as this will allow researchers to better understand the relationship between the SET and TSR, as well as the degree to which information gleaned from them overlaps.

Second, the schools included in this psychometric study had been implementing their CI3T models for a varying number of years. Five schools were in their first year of implementation, one school was in its third, one in its fourth, and one in its fifth year. Although we know the SET lacks variability regardless of how long a school has been implementing, it is unclear how sensitive the TSR is to a school's years of experience with their CI3T plan. Conceivably, TSR ratings could change over time. For example, implementation may be high in the first few years of implementation, but as the plan loses novelty, implementation may decline. Or, as schools become more adept at implementation, implementation (and hence, TSR scores) may improve. Thus, future research should involve assessment of schools implementing for the same number of years to control for variability associated with years of experience. Additionally, researchers should examine how TSR scores change over time.

Third, it is important to acknowledge the TSR relies on self-report. Although the TSR is the first such measure to assess primary plan implementation within individual classrooms, having teachers and instructional staff report their own levels of implementation may not constitute the most reliable data source. That is, self-report scores tend to be higher than those reported by outside observers (e.g., Lane, Kalberg, Bruhn, et al., 2008) making it difficult to distinguish between the true score and the self-reported score. In an effort to continue assessing implementation in individual classrooms, direct observation by unbiased observers is a logical next step to providing a

more accurate picture of classroom implementation. Further, direct observation scores and self-report scores should be compared to see exactly how scores vary by rater.

Fourth, only one estimate (i.e., internal consistency) of reliability was assessed. Although internal consistency of the TSR was good, multiple estimates (e.g., alternate-form, test-retest) provide stronger evidence regarding the reliability of a measure than a single estimate alone (AERA, APA, & NCME, 1999). As Benson (1998) pointed out, multiple data analytic techniques from the structural (e.g., reliability estimates) and external (i.e., validity estimates) stages should be used in the validation process. Because the TSR was completed only one time during the school year, adding an additional assessment point within the same school year by the same raters would allow future researchers to determine test-retest reliability (Hatcher & Stepanski, 1994) further contributing to the psychometric analysis of the TSR.

Finally, while an internal consistency estimate provides some evidence of reliability, more sophisticated data analytic techniques are available. Techniques such as principal component analysis (PCA), which require a larger sample size, are recommended when statistical assumptions (e.g., approximately equal interitem correlations when estimating subscale alpha values) are not met (Cronbach & Shavelson, 2004; Vincent et al., 2010). Although the TSR creators were well-versed in the theory, research, and practice of CI3T implementation allowing them to develop accurately a treatment integrity assessment tool addressing core CI3T components; PCA would provide additional evidence about which items actually load onto certain dimensions (Hatcher & Stepanski, 1994). It is possible primary plan implementation consists of more or less than the three dimensions—teaching, reinforcing, and monitoring—included

in the TSR. Additionally, PCA would allow for the number of items to be reduced to those accounting for the most variance, thus creating a shorter and more accurate assessment tool. For schools choosing to evaluate CI3T primary plan implementation, clearly a shorter and more accurate tool would be desirable—especially for teachers who often lack time during the instructional day. Thus, future researchers should consider conducting a PCA with a larger sample to further understand the technical adequacy of the TSR.

Conclusion

As schools move forward in their implementation of CI3T models, it is critical they evaluate their level of primary plan implementation before allocating resources to support students and teachers alike. Further, we contend assessment is needed at the classroom level because it is within classrooms that students most frequently are exposed to academic, behavioral, and social skills presumably taught and reinforced by teachers and instructional staff. Accurate classroom-level evaluation requires an assessment tool that produces reliable and valid results consistently over time. To meet this need, we developed the TSR which demonstrated initial evidence of internal consistency reliability. The TSR also demonstrated divergent evidence of validity, that is to say, it does not measure the same constructs as the SET. Although we recognize the promise of the TSR to accurately measure primary plan implementation of CI3T models at the classroom level above and beyond the information captured by the SET, further development of the TSR is needed if schools are going to use data derived from the TSR to make decisions. Namely, prior to allocating resources to (a) support teachers

implementing at low levels and (b) students not responding to teachers implementing at high levels; the data used to make these decisions must be accurate. Thus, further research and development of the TSR is needed to address the limitations of this initial validation study.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999*). Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Batsche, G., Elliot, J., Graden, J. L., Grimes, J., Kovaleski, J. F., Prasse, D., et al. (2005). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: National Association of State Directors of Special Education.

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

Benson, J. (1998). Developing a strong program of test validation: A text anxiety example. *Educational Measurement: Issues and Practice, 17*, 10-22.

Billingsley, F. F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment, 2*, 229-241.

Bradshaw, C. P., Barrett, S., & Bloom, J. (2004). *The Implementation Phases Inventory (IPI).* Baltimore: PBIS Maryland. Available from http://www.pbismaryland.org/forms.htm.

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

Bruhn, A. L., Lane, K. L., & Hirsch, S. E. (2011). *A review of secondary interventions conducted within multi-tiered models of prevention evidencing a primary behavioral plan.* Manuscript in preparation.

Cheney, D., & Walker, B. (2003a). *The BEACONS Project individual positive behavior support Self-Assessment and Program Review*. Seattle: University of Washington.

Cheney, D., & Walker, B. (2003b). *The BEACONS Project positive behavior support Leadership Team Self-Assessment and Program Review*. Seattle: University of Washington.

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

Compton, D., Fuchs, D., Fuchs, L. S. & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98,* 394-409.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L., & Shavelson, R. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391-418.

DeVellis, R. F. (2006). Classical test theory. *Medical Care, 44,* 50-59.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41,* 327-350.

Elliott, S.N., & Gresham, F.M. (2008b). *Social Skills Improvement System: Intervention guide*. Bloomington, MN: Pearson Assessments.

Fairbanks, S., Sugai, G., Guardino, D., & Lathrop, M. (2007). Response to intervention: examining classroom behavior support in second grade. *Exceptional Children, 73,* 288-310.

Gast, D. L. (2010). *Single subject research methodology in behavioral sciences.* New York: Routledge.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 60,* 294-309.

Gresham, F. M., Gansle, K., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis, 26,* 257-263.

Gresham, F. M., Van, M. B., & Cook, C. R. (2006). Social skills training for teaching replacement behaviors: remediating acquisition deficits in at risk students. *Behavioral Disorders, 32,* 363-377.

Hatcher, L., & Stepanski, E. J. (1994). *A step-by-step approach to using the SAS system for univariate and multivariate statistics*. Cary, NC: SAS Institute.

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

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179.

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

Jack, S. L., Shores, R. E., Denny, R. K., Gunter, P. L., DeBriere, T., & DePaepe, P. (1996). An analysis of the relationship of teachers' reported use of classroom management strategies on types of classroom interactions. *The Journal of Behavioral Education, 6,* 67–87.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 2*5, 251-227.

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

Lane, K. L., Bocian, K. M., MacMillan, D. L., & Gresham, F. M. (2004). Treatment integrity: An essential-but often forgotten-component of school-based interventions. *Preventing School Failure, 48,* 36-43.

Lane, K. L., Graham, S., Harris, K. R., Little, M. A., Sandmel, K., & Brindle, M. (2009). The effects of self-regulated strategy development for second-grade students with writing and behavioral difficulties. *Journal of Special Education, 41*, 1-22.

Lane, K. L., Kalberg, J. R., Bruhn, A. L., Mahoney, M. E., & Driscoll, S. A. (2008). Primary prevention programs at the elementary level: Issues of treatment integrity, systematic screening, and reinforcement. *Education and Treatment of Children, 31*, 465-494.

Lane, K. L., Kalberg, J. R., & Edwards, C. (2008). An examination of school-wide interventions with primary level efforts conducted in elementary schools: Implications for school Psychologists (pp. 253-278). In D. H. Molina (Ed.) *School Psychology: 21st Century Issues and Challenges.* New York, NY: Nova Science Publishers.

Lane, K. L., Kalberg, J. R., & Menzies, H. M. (2009). *Developing schoolwide programs to prevent and manage problem behaviors: A -step-by-step approach*. New York, N.Y.: Guilford Press.

Lane, K. L., Menzies, H., & Kalberg, J. R. (in press). An integrated, comprehensive three-tier model to meet students' academic, behavioral, and social needs. In K. Harris, T. Urdan, and S. Graham (Eds.). *American Psychological Association. Educational Psychology Handbook.* Washington, DC: American Psychological Association.

Lane, K. L., Menzies, H. M, Oakes, W. P., & Kalberg, J. R. (2011). *Systematic screenings of behavior to support instruction: From preschool to high school.* Book under contract with Guilford.

Lane, K. L., Robertson, E. J., & Graham-Bailey, M. A. L. (2006). An examination of schoolwide interventions with primary level efforts conducted in secondary schools: Methodological considerations. In T. E. Scruggs & M.A. Mastropieri (Eds.), *Applications of research methodology: Advances in learning and behavioral disabilities: Vol.19.* Oxford, UK: Elsevier.

Lane, K. L., Robertson, E. J., & Wehby, J. H. (2002). *Primary Intervention Rating Scale.* Unpublished rating scale.

Lane, K. L., Wehby, J. H., Robertson, E. J., & Rogers, L. A. (2007). How do different types of high school students respond to schoolwide positive behavior support programs? Characteristics and responsiveness of teacher-identified students. *Journal of Emotional and Behavioral Disorders, 15*, 3-20.

Martson, D. (2005). Tiers of intervention in responsiveness to intervention: Prevention outcomes and learning disabilities identification patterns. *Journal of Learning Disabilities, 38*, 539-544.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the journal of applied behavior analysis. *Journal of Applied Behavior Analysis, 30*, 659-672.

Moncher, F. J., & Prinz, F. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review, 11*, 247-266.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Robertson, E. J., & Lane, K. L. (2007). Supporting middle school students with academic and behavioral concerns: A methodological illustration for conducting secondary interventions within three-tiered models of support. *Behavioral Disorders, 33*, 5-22.

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Interdisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 39*, 460-475.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99-103.

Sugai, G. & Horner, R. H. (2006) A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychology Review, 35,* 245-259.

Sugai, G., Lewis-Palmer, T., Todd, A., & Horner, R. H. (2001). School-wide evaluation tool. Eugene: University of Oregon.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2006). 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. (2001). *Team Implementation Checklists* (Version 2.2). Eugene: University of Oregon, OSEP Center for Positive Behavioral Supports.

Sutherland, K. & Oswald, D. (2005). The relationship between teacher and student behavior in classrooms for students with emotional and behavioral disorders: Transactional processes. *Journal of Child and Family Studies, 14,* 1-14.

Vincent, C., Spaulding, S., & Tobin, T. (2010). A reexamination of the psychometric properties of the school-wide evaluation tool (SET). *Journal of Positive Behavior Interventions, 12,* 161-179.

Walker, L. S., Beck, J. E., Garber, J., & Lambert, W. (2009). Children's Somatization Inventory: Psychometric properties of the revised form (CSI-24). *Journal of Pediatric Psychology, 34,* 430-440.

Walker, B., Cheney, D. & Stage, S. (2009). The validity and reliability of the self-assessment and program review: Assessing school progress in schoolwide positive behavior support. *Journal of Positive Behavior Interventions, 11,* 94-109.

Wanzek, J., & Vaughn, S. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. *Journal of Learning Disabilities, 41,* 12-142.

Wolery, M., & Ledford, J. R. (2011). *Assessing procedural fidelity in single case experiments: Rationale, methods, and reporting*. Manuscript in preparation.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156-167.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation,30,* 44-61.