Hammill Institute on Disabilities

Journal of Positive Behavior Interventions
2017, Vol. 19(4) 195–204

DOI: 10.1177/1098300717693384
jpbi.sagepub.com
SAGE

Comparability of Fidelity Measures for Assessing Tier 1 School-Wide Positive Behavioral Interventions and Supports

Sterett H. Mercer, PhD[1], Kent McIntosh, PhD[2], and Robert Hoselton, BS[2]

Abstract

Several reliable and valid fidelity surveys are commonly used to assess Tier 1 implementation in School-Wide Positive Behavioral Interventions and Supports (SWPBIS); however, differences across surveys complicate consequential decisions regarding school implementation status when multiple measures are compared. To address this concern, the current study (a) provides updated convergent validity estimates for five fidelity measures, (b) tests mean differences in scores and reports the percentages of schools meeting recommended implementation criteria by measure, and (c) investigates sensitivity of the measures to differences between schools at varied levels of implementation. Across most surveys, convergent validity estimates were moderate (r = .59–.71), and mean differences were negligible for all surveys other than the School-Wide Evaluation Tool (SET), on which higher scores were more likely to be obtained. Despite higher average scores, the SET classified similar percentages of schools as adequately implementing compared with other measures with a 70% implementation criterion, but fewer schools when compared with measures with an 80% criterion. Compared with other measures, the PBIS Self-Assessment Survey (SAS) was more sensitive to differences among schools at higher levels of implementation. Implications for SWPBIS research and fidelity assessment are discussed.

Keywords

school-wide positive behavioral interventions and supports, fidelity, implementation

School-Wide Positive Behavioral Interventions and Supports (SWPBIS) is a multitiered framework to improve student behavior and academic performance that is currently implemented in more than 23,000 schools (Horner, 2016). At the universal Tier 1 level, core features of SWPBIS include (a) staff commitment to a formal, proactive approach to school discipline, (b) identification of a small set of positively stated behavioral expectations, (c) teaching of these behavioral expectations across school settings, (d) a system for acknowledging or rewarding students who display these behavioral expectations, (e) specification of a range of consequences for problem behavior, and (f) use of data for decision making (Sugai & Horner, 2009). One form of data that has been emphasized in SWPBIS is fidelity of implementation, which is defined as the extent to which an intervention is delivered as intended (Domitrovich et al., 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003). This focus on fidelity of implementation is warranted because higher SWPBIS implementation fidelity is associated with better student outcomes. For example, in a SWPBIS high school trial, Flannery, Fenning, Kato, and McIntosh (2014) found that as fidelity of implementation increased, exclusionary discipline significantly decreased.

Fidelity of implementation is often conceptualized by researchers primarily as an external evaluation of the extent to which schools are implementing a practice adequately enough to be classified as such for research or evaluation purposes. In addition, schools implementing SWPBIS are encouraged to self-assess fidelity, for the purposes of monitoring implementation over time and identifying next steps for implementation. Routine assessment of implementation fidelity is useful because team use of data for decision making improves the likelihood of sustained SWPBIS implementation (Coffey & Horner, 2012; McIntosh, Kim, Mercer, Strickland-Cohen, & Horner, 2015; McIntosh et al., 2013).

To facilitate data-based decision making, a variety of reliable and valid SWPBIS fidelity surveys have been developed and are widely used, particularly to assess Tier 1 implementation features. Although they assess similar core

[1]The University of British Columbia, Vancouver, Canada
[2]University of Oregon, Eugene, USA

Corresponding Author:
Sterett H. Mercer, University of British Columbia, 2125 Main Mall, Vancouver, British Columbia, Canada, V6T 1Z4.
Email: sterett.mercer@ubc.ca

Action Editor: Brandi Simonsen

Table 1. Key Features of Tier 1 Fidelity Measures.

Measure	Number of items	Intended use	Respondents	Response process	Frequency	Implementation criteria
School-Wide Evaluation Tool (SET) Version 2.1	28	Annual evaluation	External evaluator	Evaluator completes based on student and staff interviews, school walkthrough, and permanent product review	Annually	80% on both Behavioral Expectations Taught subscale and total
School-Wide Benchmarks of Quality (BOQ) Revised	53	Annual evaluation and action planning	SWPBIS team members and internal or external coach	SWPBIS team and coach complete independently, then coach aggregates into final scores	Annually	70% on total
PBIS Self-Assessment Survey (SAS) Version 2.0	18 (*School-Wide Systems* scale)	Determine staff perceptions on fidelity	All school staff members	Individual staff responses are averaged by school	Annually to triennially	80% on School-Wide Systems Implementation Average
Team Implementation Checklist (TIC) Version 3.1	22	Progress monitoring and action planning during initial implementation	SWPBIS team members	SWPBIS team completes collaboratively	3–4 times per year	80% on total
SWPBIS Tiered Fidelity Inventory (TFI)	15 (Tier 1 scale)	Progress monitoring, annual evaluation, and action planning	SWPBIS team members with external coach	Team and coach complete collaboratively based on student and staff interviews, school walkthrough, and permanent product review	At least annually	70% on Tier 1 scale

Note. SWPBIS = School-Wide Positive Behavioral Interventions and Supports.

Tier 1 features of SWPBIS, the surveys vary in terms of type (i.e., internal self-assessment or external evaluation), response process (i.e., completed by one person, collaboratively by multiple members, or by multiple staff members), intended frequency of use (i.e., multiple times per year for progress monitoring versus annually), and recommended criteria (i.e., cut scores) for adequate implementation. These aspects of the Tier 1 surveys are summarized in Table 1. Due to these differences, it can be difficult to compare scores across measures, to anticipate what an obtained score might be on a specific measure that was not completed, and to determine whether schools using different fidelity measures are adequately implementing core features of Tier 1 SWPBIS.

These issues of score comparability can be conceptualized as concerns related to construct validity (Messick, 1995). Users of these measures (e.g., schools, districts, state SWPBIS networks) make consequential judgments about the extent to which schools are adequately implementing critical Tier 1 SWPBIS features to inform action planning, resource allocation, annual district and state SWPBIS evaluations, and public recognition of schools and districts with high levels of implementation and associated student outcomes (e.g., Barrett, Bradshaw, & Lewis-Palmer, 2008; Bradshaw et al., 2012; Horner et al., 2014; Simonsen et al., 2012; Upreti, Liaupsin, & Koonce, 2010), and these consequential decisions may be inappropriate if scores and implementation criteria are not comparable across fidelity measures.

The scores may appear to be comparable with users because the measures sample similar content domains (e.g., the extent to which positively stated behavioral expectations are identified and taught), scores are reported similarly across measures (i.e., percentage of total points or features being implemented adequately), and each measure has an expert-recommended criterion for adequate implementation that can be used to classify schools as adequately (or not adequately) implementing SWPBIS. Table 2 (developed by the authors) provides a cross-reference of item content domains by fidelity measure.

Despite these similarities, the fidelity measures differ in key ways, such as (a) the specificity with which core SWPBIS elements are assessed, (b) the relative weighting of these elements in the calculation of total scores due to

Table 2. Item Content Cross-Reference for Tiered Fidelity Inventory With Other Tier 1 Fidelity Measures.

TFI Tier 1 item	SET item	BOQ item	SAS item	TIC item
1.1 Team Composition	F2, F2, F3, F4, F5	1	9, 10	1, 3, 20
1.2 Team Operation Procedures	F6, F8	2, 3		4, 8, 16
1.3 Behavioral Expectations	A1, A2	17, 18, 19, 20	1	9, 10
1.4 Teaching Expectations	B1, B2, B4	29, 30, 31, 32, 33, 37, 38, 50	2, 14	11, 12
1.5 Problem Behavior Definitions	D2	7, 8, 10, 11	4, 5, 6	14
1.6 Discipline Policies	D1	8, 12, 51	7, 8	14
1.7 Professional Development	B3	35, 36, 38, 40	17	16
1.8 Classroom Procedures		42, 43, 44, 45, 46, 47, 48		15
1.9 Feedback and Acknowledgment	C1, C2, C3	22, 23, 24, 25, 26, 52	3	13
1.10 Faculty Involvement	E3, F7	4, 5, 6, 16, 21, 33, 49	12, 16	2, 11, 18
1.11 Student/Family/Community Involvement		27, 33, 34, 41, 49	13	
1.12 Discipline Data	E1, E2	8, 9, 13	11	17, 18, 19
1.13 Data-based Decision Making	E4	15, 16, 38, 53	12, 14	7, 8, 19
1.14 Fidelity Data				6, 15, 16
1.15 Annual Evaluation		14, 53	18	

Note. All of these measures are available for download at no cost at https://www.pbisapps.org/Applications/Pages/PBIS-Assessment-Surveys.aspx. SAS item numbers are for the School-Wide Systems subscale. Due to differences across the measures, some items have multiple cross-references and others have none. TFI = Tiered Fidelity Inventory; SET = School-Wide Evaluation Tool; BOQ = School-Wide Benchmarks of Quality; SAS = PBIS Self-Assessment Survey; TIC = Team Implementation Checklist; SWPBIS = School-Wide Positive Behavioral Interventions and Supports.

differences in number of items assessing each domain, (c) the response process used to complete the measures, and (d) the specific cut scores used to determine whether a school is adequately implementing SWPBIS. For these reasons, greater attention to the convergent validity of these measures is needed to evaluate the appropriateness of the consequential decisions made based on these fidelity scores.

These consequential decisions made by practitioners rely on different aspects of measure comparability. For example, district or state teams may allow school teams to choose from a range of self-assessment tools for their action planning. If these measures assess different aspects of fidelity (i.e., they have low convergent validity), the scores will indicate different items for action planning, which would complicate team goals and district technical assistance efforts. An even more important consequential decision is determining whether schools are implementing SWPBIS to criterion when different fidelity measures are used. Evaluators may want to answer this question using a range of existing measures instead of requiring all schools to use the same tool. If the criteria for adequate implementation are not comparable, the resulting classifications will not be accurate (i.e., dependent on the measure used in addition to the level of implementation). For example, there are distinct negative consequences for false positives (i.e., a school's inadequate implementation is misclassified as adequate) and false negatives (i.e., a school's adequate implementation is misclassified as inadequate). Similarly, school, district, and state teams benefit from measures that produce predictable scores not just for implementation classification but also for assessing growth in fidelity. Some measures may have floor or ceiling effects that differentially affect scores at varied levels of implementation. If so, different measures may not accurately reflect improvements in implementation over time.

Findings regarding convergent validity would also be helpful in future research on factors predicting sustained implementation and on the association of SWPBIS fidelity and student outcomes given that inconsistent use of SWPBIS fidelity assessments by schools has been a complication in these studies (e.g., McIntosh et al., 2013; Pas & Bradshaw, 2012; Simonsen et al., 2012).

In an attempt to address these concerns regarding comparability of Tier 1 fidelity assessments, we investigated the following research questions:

Research Question 1: To what extent is there evidence of convergent validity across Tier 1 SWPBIS measures? This question addresses the extent to which the measures assess the related construct of fidelity of SWPBIS Tier 1 implementation. Moderate to high associations would indicate that the measures are assessing the same construct.

Research Question 2: Are there mean differences in scores across fidelity measures that could complicate judgments regarding whether or not a school is adequately implementing SWPBIS? This question addresses the extent to which the measures generate comparable

Table 3. Descriptive Information and Tests of Mean Differences for Pairs of Fidelity Assessments.

Measure 1	Measure 2	*n*	%Fid. 1	%Fid. 2	M_1	M_2	Difference	SD_1	SD_2	*d*	*r*
SET	BOQ	1,103	78	78	89.34	79.60	9.74***	11.78	16.27	.69	.63***
	SAS	2,055	75	49	87.23	77.08	10.15***	15.03	13.10	.72	.64***
	TIC	1,269	72	58	86.53	78.58	7.95***	14.00	15.79	.53	.59***
	TFI	36	61	58	73.89	62.03	−11.86***	28.20	32.31	.39	.92***
BOQ	SAS	3,705	82	56	81.17	79.20	1.97***	15.68	11.77	.14	.68***
	TIC	1,553	76	65	78.45	80.54	2.10***	18.39	16.89	.12	.71***
	TFI	200	90	90	84.83	85.39	0.57	13.87	15.11	.04	.65***
SAS	TIC	3,706	39	48	73.36	74.02	−0.63**	14.57	19.17	.04	.67***
	TFI	613	60	80	79.66	80.70	1.04	12.48	18.44	.10	.70***
TIC	TFI	119	56	76	75.87	77.03	1.16	21.84	21.74	.05	.96***

Note. % Fid. = percentage of schools in the sample at or above the fidelity criterion for the measure (see Tables 1 or 4 for criterion values). SET = School-Wide Evaluation Tool; BOQ = School-Wide Benchmarks of Quality; SAS = PBIS Self-Assessment Survey; TIC = Team Implementation Checklist; TFI = Tiered Fidelity Inventory.
**$p < .01$. ***$p < .001$.

scores and summative decisions regarding adequate implementation.

Research Question 3: To what extent are the fidelity measures sensitive to variability in SWPBIS implementation below, near, and above recommended cut scores for adequate fidelity on the measures? This question addresses the extent to which scores on the measures vary at different levels of implementation.

Method

Sample

To determine eligible schools and fidelity assessments, we first selected the last fidelity assessment of the school year (2005–2006 through 2014–2015) for each of five measures for schools that reported fidelity data to the Office of Special Education Programs (OSEP) National Technical Assistance Center on PBIS. For each pair of fidelity measures, we then selected schools that had completed both fidelity assessments within 30 days in the same academic year. Thus, the sample for each pairwise comparison of fidelity measures differed (see Table 3), and school demographic information also differed to some extent across each of the 10 fidelity measure comparisons. The following characteristics, based on schools that completed at least one PBIS fidelity measure in 2012–2013, are representative of schools reporting data to the OSEP National Technical Assistance Center on PBIS. Characteristics for these schools are presented for 1 year only because some schools were included in samples across multiple years. The schools were located in urban (31%), suburban (29%), town (14%), or rural (26%) areas in 42 different U.S. states. Regarding grade levels served, 67% were elementary schools, 20% were middle schools, and 11% were high schools. Of these schools, 72% were eligible for Title I programs based on student economic need. On average, 51% of students in the schools were eligible for free or reduced price meals (*SD* = 25%), and the average school racial and ethnic student composition was the following: 57% White (*SD* = 33%), 18% Hispanic or Latino (*SD* = 24%), 18% Black or African American (*SD* = 25%), 4% Asian (*SD* = 7%), 3% Two or more races (*SD* = 4%), and less than 1% American Indian or Alaskan Native and Native Hawaiian or Other Pacific Islander.

Measures

School-Wide Evaluation Tool (SET). The SET 2.1 (Sugai, Horner, & Lewis-Palmer, 2001), is a 28-item external assessment of Tier 1 SWPBIS practices that is typically completed annually based on staff and student interviews, a school observation, and permanent product reviews. Because it is completed by external evaluators, it has been viewed as the most objective and direct assessment of implementation, and therefore less influenced by individual perceptions or ratings members with varying experience with SWPBIS. Calculation of ordinal alpha (Zumbo, Gadermann, & Zeisser, 2007) based on all available SET assessments from 2010 to 2015 ($n = 10{,}640$) provides evidence of reliability for the total score (ordinal $\alpha = .95$) and the Behavioral Expectations Taught subscale (five items; ordinal $\alpha = .90$). Schools are considered to be implementing Tier 1 adequately when both the SET total and Behavioral Expectations Taught subscale scores are at or above 80% because "change in student behavior is unlikely before a school teaches the school-wide expectations and that stability of the effect is unlikely without the constellation of practices in the remainder of the SET" (Horner et al., 2004, p. 11).

School-Wide Benchmarks of Quality (BOQ). The BOQ Revised (Kincaid, Childs, & George, 2010) is a 53-item Tier 1 annual evaluation that is a combination of SWPBIS

team members' perspectives and the perspective of an internal or external coach. The team and coach either fill out the BOQ collaboratively, or the team members provide individual ratings that are consolidated into a final rating by the coach. Calculation of ordinal alpha based on all available BOQ assessments from 2010 to 2015 (n = 20,109) provides evidence of reliability for the total score (ordinal α = .98). Previous research has shown that BOQ scores are moderately correlated with the SET (r = .51), providing some evidence of convergent validity (Cohen, Kincaid, & Childs, 2007). Schools evaluated at or above 70% of the total points on the BOQ are considered to be adequately implementing Tier 1.

PBIS Self-Assessment Survey (SAS). The SAS 2.0 (Sugai, Horner, & Todd, 2000) is a 43-item annual, internal evaluation of Tiers 1 to 3 that can be completed by all staff members in a school. The SAS is intended both as a survey of fidelity of specific SWPBIS features for a broad sample of staff beyond just the team and a needs assessment for next steps in implementation. For the current study, the 18-item School-Wide Systems Implementation Average score was used, which is a measure of average staff perceptions on what proportion of Tier 1 elements are being implemented. Calculation of coefficient alpha based on all available SAS assessments from 2019 to 2015 (n = 38,362) provides evidence of reliability for the School-Wide Systems Implementation Average score (α = .97), and there is some evidence of convergent validity with the SET (.75; Horner et al., 2004). Schools scoring at or above 80% on the Implementation Average are considered to be adequately implementing Tier 1.

Team Implementation Checklist (TIC). The TIC 3.1 (Sugai et al., 2001) is a 22-item internal evaluation of Tier 1 features that is used as a progress monitoring measure (3–4 times per year) by school SWPBIS teams during initial implementation. The measure is primarily useful in guiding teams through the typical startup activities of exploring and installing SWPBIS. Calculation of ordinal alpha based on all available TIC assessments from 2009 to 2015 (n = 18,346) provides evidence of reliability for the total score (ordinal α = .95), and a confirmatory factor analysis supported its factor structure (McIntosh, Mercer, Nese, Strickland-Cohen, & Hoselton, 2016). Schools obtaining scores at or above 80% are considered to be implementing Tier 1 adequately.

SWPBIS Tiered Fidelity Inventory (TFI). The TFI (Algozzine et al., 2014), the most recently developed SWPBIS fidelity measure, allows separate assessments of the three tiers of SWPBIS in one instrument, with separate scale scores for each tier and the option of an overall implementation score. It is intended to be completed by the school SWPBIS team, with facilitation from an external coach or coordinator who is knowledgeable about SWPBIS systems. It also includes a glossary of key terms used, which can help teams that are self-assessing their implementation improve understanding of the items that they are scoring. The Tier 1 scale includes 15 items. Calculation of ordinal alpha based on all available TFI assessments from 2014 to 2015 (n = 2,160) provides evidence of reliability for the Tier 1 score (ordinal α = .97), and prior technical adequacy studies (Massar, McIntosh, & Mercer, 2017; McIntosh et al., 2017) provide evidence of content validity, factor structure, and reliability (Tier 1 coefficient α = .87; interrater and 2-week test-retest intra class correlations = .99), as well as evidence of convergent validity with other Tier 1 SWPBIS fidelity measures (r = .54–.64). Schools are considered to be implementing Tier 1 adequately when scores are at or above 70%.

Procedures

Extant data were retrieved from PBIS Assessment, a database housed at the University of Oregon (http://pbisapps.org), a free online application system available to any U.S. school so long as they have (a) an identified PBIS coordinator and (b) agree to the use of their data for research purposes. PBIS Assessment allows users to enter and track fidelity of implementation and outcome data to facilitate data-based decision making and evaluation. The SWPBIS fidelity data used in this study were entered by school or district personnel into PBIS Assessment and extracted by the research team; thus, there was variability in the specific roles of respondents within and across fidelity measures, and data on fidelity of assessment procedures are not available.

Data Analyses

To address Research Question 1 regarding convergent validity, we calculated Pearson's r for each pair of fidelity assessments. To address Research Question 2 regarding the presence of mean differences on fidelity assessments that could complicate judgments of whether or not a school is adequately implementing, we conducted paired sample t tests for each pair of measures and calculated standardized mean difference (d) effect sizes using the following formula (Dunlap, Cortina, Vaslow, & Burke, 1996):

$$d = \frac{\left(M_1 - M_2\right)}{SD},$$

where M_1 and M_2 are mean scores on fidelity assessments, and SD is the pooled standard deviation for the two fidelity assessments. To further address this research question, we compared the percentage of schools that were at or above the recommended fidelity criteria for the measures in the samples of paired fidelity assessments.

To address Research Question 3 regarding the extent to which the fidelity assessments are sensitive to variability in implementation below, near, and above recommended criteria for implementation status, we conducted equipercentile score linking with log-linear polynomial presmoothing (see Kolen & Brennan, 2004). In equipercentile linking, each school's score on a target fidelity assessment is transformed such that the score has the same percentile rank as on a reference fidelity assessment (i.e., the linked scores have the same percentile rank relative to the group of schools). We selected equipercentile linking instead of other linear linking measures because we did not expect the relations among scores on different fidelity assessments to be consistent across the distribution of scores on the measures. Because scores in samples tend to be less smooth (i.e., score frequencies change more erratically across the score distribution) than scores in populations, it is customary to smooth scores before linking to reduce sample-specific irregularities. Based on comparisons of model fit for each pair of assessments, we used bivariate log-linear smoothing models that included polynomial terms up to the fourth power (thus preserving the mean, variance, skew, and kurtosis) and a two-way interaction term. Because observed score linking methods require large sample sizes, TFI scores were linked only with the SAS (n = 613); sample sizes for linking with other Tier 1 measures were all large (minimum n = 1,103).

Results

Table 3 presents sample sizes for schools that completed pairs of fidelity assessments within 30 days, the percentage of schools meeting the recommended criteria for adequate implementation and the means and standard deviations on each assessment in the pair, tests of mean differences based on paired sample t tests, and Pearson correlations for pairs of assessments.

Convergent Validity and Mean Differences

In general, convergent validity among all assessments was moderate, with rs ranging from .59 to .71 ($p < .001$), with the exception of two comparisons with the smallest sample sizes, the SET with the TFI ($r = .92$, $p < .001$, $n = 36$), and the TIC with the TFI ($r = .96$, $p < .001$, $n = 119$). Regarding mean differences in scores, scores on the SET were consistently higher compared with all other fidelity assessments (by 7.95–11.86 percentage points, all $p < .001$, $d = .39$–.72). Encouragingly, there were no statistically significant mean differences between the TFI and the BOQ, SAS, or TIC (d = .04–.10). Although there were statistically significant mean differences between scores on the BOQ, SAS, and TIC, most likely due to the very large sample sizes for these comparisons (smallest n = 1,553), the actual mean differences were of negligible to small magnitude (0.63–2.10 percentage points, d = .04–.14).

The higher scores obtained on the SET did not consistently manifest in more schools being classified as at or above the criteria for adequate implementation, however, because the recommended criteria for adequate implementation on the SET includes two scores (≥80% on both the total score and the Behavior Expectations Taught subscale, known as the 80–80 criterion). For fidelity assessments with a criteria of ≥70% on the total score (i.e., BOQ and TFI) and the SET with the 80–80 criterion, approximately similar percentages of schools were classified as adequately implementing. Specifically, 78% of schools were classified as adequately implementing on both assessments in the sample comparing the SET and BOQ (n = 1,103); 61% and 58% as adequately implementing on the SET and TFI, respectively (n = 36); and 90% adequately implementing on both the BOQ and TFI (n = 200). By contrast, more schools were consistently classified as adequately implementing on the SET, BOQ, and TFI when compared with assessments with greater than 80% criteria (i.e., the SAS and TIC). For example, 75% versus 49% of schools were classified as adequately implementing in the comparison of the SET and SAS (n = 2,055), and 72% versus 58% of schools were classified as adequately implementing when comparing the SET and TIC (n = 1,269).

Score Linking

Equipercentile linked scores for the measures are presented graphically in Figure 1. In addition, linked scores on each fidelity measure at the recommended implementation criterion (based on Tier 1 total scores) for each measure are presented in Table 4. If linked scores on one fidelity measure were perfectly equivalent to another measure, data points would be on the diagonal (dotted line); if higher scores on the second measure (as listed in the figure legend) are more likely to be obtained than on the first measure, data points are above the diagonal, with the opposite pattern (below diagonal) if higher scores are likely to be obtained on the first measure. Curved lines in the figure indicate that score differences between measures are not constant across levels of implementation (i.e., differences are more pronounced at particular levels of implementation). In the figure, linked scores for the comparisons involving the SET (as the second measure) were consistently above the diagonal (i.e., schools administered the SET and other fidelity measures had consistently higher SET scores), reflecting the higher scores on the SET that were reported in the tests of mean differences and indicating less sensitivity of the SET to capture differences in higher SWPBIS implementation levels relative to other fidelity measures.

For SAS scores compared with all other measures, there is asymptotic curvature at the upper end of the linked score

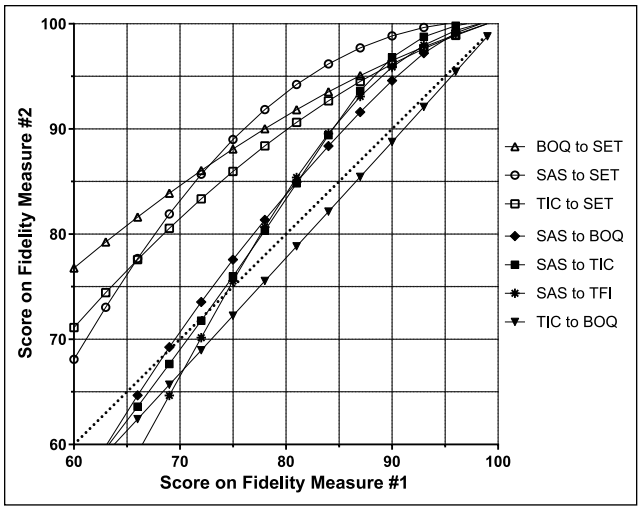

Figure 1. Linked scores on pairs of fidelity assessments.
Note. The dotted line in the figure serves as a reference for equivalent linked values on fidelity assessments. Sample sizes are reported in Table 3. BOQ = School-Wide Benchmarks of Quality; SET = School-Wide Evaluation Tool; SAS = Self-Assessment Survey; TIC = Team Implementation Checklist; TFI = Tiered Fidelity Inventory.

Table 4. Linked Scores at Criteria for Adequate Implementation Fidelity.

Measure 1	Recommended criterion	Measure 2	Linked score
SET[a]	80	BOQ	64
		SAS	68
		TIC	68
BOQ	70	SET	85
		SAS	70
		TIC	73
SAS	80	TFI	84
		SET	93
		BOQ	84
		TIC	83
TIC	80	SET	90
		BOQ	78
		SAS	78
TFI	70	SAS	72

Note. SET = School-Wide Evaluation Tool; BOQ = School-Wide Benchmarks of Quality; SAS = PBIS Self-Assessment Survey; TIC = Team Implementation Checklist; TFI = Tiered Fidelity Inventory.
[a]The recommended criterion for the SET for adequate implementation is a score of at least 80 on the total and Behavioral Expectations Taught subscale—the linking analyses are based only on the SET total score.

distributions, indicating that there is less of a score ceiling on the SAS compared with other measures (including the TFI) and thus more sensitivity of the SAS to differences in higher levels of SWPBIS implementation compared with other measures. For measures other than the SET (which tends to yield higher scores across the distribution), higher scores were more likely to be obtained on the SAS than on other measures when scores were roughly below 70%, indicating less sensitivity to differences in lower levels of SWPBIS implementation. When scores were approximately in the range of 70% to 80%, the linked scores of the SAS with the TFI, BOQ, and TIC were approximately equal, and then lower scores were likely to be obtained on the SAS when above 80%, in part reflecting the score ceiling on the other measures described previously (i.e., SAS scores were consistently lower when the percent implementation was high). By contrast, the relation of TIC and BOQ scores was more consistent across the score distribution, as indicated by the line with very little curvature in Figure 1, with higher scores more likely to be obtained on the TIC.

Discussion

The purpose of this study was to examine the comparability of SWPBIS Tier 1 fidelity assessments to inform consequential judgments about SWPBIS implementation quality through (a) updated convergent validity estimates, (b) tests of mean differences in obtained scores and comparisons of percentages of schools that would be classified as adequately implementing across measures, and (c) score linking analyses to examine the sensitivity of the measures to differences in at varied levels of implementation. For all assessments, convergent validity estimates were moderate (r = .59–.71) aside from two estimates including the TFI with smaller sample sizes (r = .92 and .96). Although convergent validity estimates for the older Tier 1 measures have been reported in prior studies (e.g., Horner et al., 2004), the updated estimates in the current study are based on larger samples than in prior studies (1,103–3,706 schools) and required the measures to be completed within 30 days. As a result, readers can view these findings as more representative and accurate tests of convergent validity. The TFI sample sizes are smaller because fewer years of data are available; however, the reported validity estimates are higher than in preliminary studies of the TFI (McIntosh et al., 2017), possibly due to the more stringent 30-day administration time window used in the current study, which unfortunately also reduced the number of paired assessments available for analyses involving the TFI.

Regarding mean differences across assessments, the primary finding is that total scores on the SET were significantly higher than on all other Tier 1 fidelity assessments (7.95–11.86 percentage points, d = .39–.72), indicating that higher total scores are more likely to be obtained on the SET than on other measures for similar levels of Tier 1 implementation. For other fidelity assessments, there were few mean differences in scores. Of note, there were no statistically significant differences between the Tier 1 scores of the TFI and the BOQ, SAS, and TIC (d = .04–.10). For the BOQ, SAS, and TIC, there were statistically significant

differences, in part due to very large sample sizes, but the actual mean differences were of negligible to small magnitude (0.63–2.10 percentage points, $d = .04–.10$).

Despite the higher average total scores on the SET, the SET dual implementation criteria (i.e., at least 80 on the total and Behavioral Expectations Taught subscale) appeared to reduce the potential effect of higher average scores on the percentages of schools classified as adequately implementing Tier 1 SWPBIS. Specifically, the SET had roughly similar classification rates to the measures with 70% or greater criteria for adequately implementing (i.e., the BOQ and TFI), and these three measures classified more schools as adequately implementing than the measures with 80% or greater criteria (i.e., the SAS and TIC). Consequently, although the tests of mean differences indicated trivial differences in the total scores across measures, other than the consistently higher scores on the SET compared with other measures, differences in the recommended criteria for adequate implementation across the measures could lead to substantive differences in decisions about whether or not a school is adequately implementing SWPBIS.

The score linking analyses present a more nuanced perspective on the differences among these fidelity assessments, particularly for the SAS in relation to the other measures (TFI, BOQ, and TIC). When implementation levels are low (i.e., <70%), SAS scores are likely to be higher than other measures, and thus the SAS may be less sensitive to differences among schools at lower levels of SWPBIS implementation. When implementation levels are within or near the range of adequate implementation (between approximately 70% and 80%), SAS scores are roughly equivalent to scores on the other measures. As implementation levels move above this range, however, SAS scores are likely to be lower than on the other measures because the SAS is less affected by score ceilings and thus may be more sensitive to differences among schools at higher levels of SWPBIS implementation. For these higher levels of implementation, this finding is not surprising—unlike other fidelity assessments, the invitation for all staff members to complete surveys that are then aggregated on the SAS requires both (a) high levels of implementation and (b) consensus among staff members that these implementation features are in place for high scores to be obtained. By contrast, the small mean difference in scores on the BOQ compared with the TIC was more consistent across levels of SWPBIS implementation.

Limitations and Future Research

The primary limitation of the study is that the measures were collected through an extant database, and therefore, the exact composition of the group completing the measures and level of adherence to administration rules are both unknown. It is likely that except for the SET and SAS, a similar group of individuals completed both pairs of assessments. This possibility may have inflated convergent validity estimates. It also indicates the need for future research to examine how respondent membership (e.g., with vs. without coach, with vs. without administrator) affects behavior during the assessment and scoring, and it would also be helpful to examine the consistency of these responses to objective observations of SWPBIS implementation. Finally, the ceiling effects seen for measures other than the SAS indicate the need to develop tools to assess fidelity of implementation for the highest levels of SWPBIS implementation.

Implications for Research

Because all of the SWPBIS measures are publicly available and familiar to implementers, it is likely that researchers will encounter schools using some combination of these measures in their prospective samples. The correlations and tests of mean differences show that the measures are related to one another and that the total scores can be used similarly to indicate level of implementation, although there appear to be substantive differences in classifications of whether a school is implementing adequately depending on the recommended criterion of the measure. This study provided evidence that the current criteria for some measures (BOQ, SET, and TFI) may be less stringent than others (SAS and TIC). These differences in classification rates have implications for a common research practice for assessing fidelity status across schools using different measures, i.e., using the implementation criteria for each measure (see Table 1) to indicate whether schools are implementing SWPBIS adequately (McIntosh et al., 2013; Nese et al., 2016) because fewer schools may be classified as adequately implementing on the SAS and TIC compared with the other measures. By contrast, the moderate correlations among total scores found in the current study are consistent with a recent study on a separate sample showing that these measures all load onto a single latent factor of Tier 1 SWPBIS fidelity of implementation with strong model fit (Turri et al., 2016). These findings add to the evidence that the total scores from these measures may be used relatively interchangeably (with the caveats described above) for research purposes, but researchers should be cautious about the application of implementation criteria to classify schools as adequately implementing or not when multiple measures are used or compared. Future research should continue to examine the empirical foundations for these criteria against alternate cut points.

Implications for Practice

Practitioners and evaluators may use these findings to see how scores on one measure may relate to scores on other measures, particularly for districts or states that have more variability in fidelity assessment use, or those considering the switch to newer measures such as the TFI. However, it

is important to stress that the analyses we conducted are not true equivalence tests due to differences in item content and coverage of critical aspects of SWPBIS implementation, and using these results for score conversions would be inappropriate. The measures are not identical, and each was designed for a specific purpose. The difference between a self-assessment checklist for initial implementation (TIC), a survey and needs assessment for the entire staff (SAS), and an external evaluation (SET) should be obvious to experienced users and evaluators. As such, it would be counterproductive to try to select measures based on how easy it may be to obtain a minimum score for adequate implementation. As described previously, the key benefits from fidelity assessment come from their use as assessments of next steps for implementation, not simply meeting a criterion and then halting further implementation efforts. As a result, practitioners are advised not to rely too heavily on these criteria as precise discriminators between adequate and inadequate SWPBIS implementation. In addition, relying only on self-reported fidelity of implementation for district or state recognition systems seems unwise, as any potential bias toward inflated scores would reduce their utility for action planning. Instead, state and district teams are encouraged to use improvements in student outcomes (e.g., reduced use of suspensions, improved perceptions of school climate, equity in school discipline) as criteria.

Many district and state evaluators have relied on the SET as an objective assessment of implementation, but these results indicate that SET scores are generally higher than scores on the other measures, with a ceiling that may not provide teams with multiple years of suggested action planning items. Instead, those seeking a balance between external evaluation and utility for implementation action planning could consider using measures like the TFI and BOQ with an external coach or technical assistance provider present to ensure that responses are accurate (McIntosh et al., 2017). Such an assessment plan, especially when paired with a periodic (every few years) assessment of all staff perceptions with the SAS, may have the most promise for assessing high-quality implementation.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A120278 to University of Oregon.

References

Algozzine, R. F., Barrett, S., Eber, L., George, H., Horner, R. H., Lewis, T. J., . . . Sugai, G. (2014). *SWPBIS Tiered Fidelity Inventory*. Eugene, OR: OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports. Available from http://www.pbis.org

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

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

Coffey, J., & Horner, R. H. (2012). The sustainability of school-wide positive behavioral interventions and supports. *Exceptional Children*, *78*, 407–422.

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.

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., . . . Ialongo, N. S. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion*, *1*, 6–28.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. doi:10.1037/1082-989X.1.2.170

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

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

Horner, R. H. (2016, August). *Schools implementing PBIS* (Technical working paper).Eugene, OR: OSEP Center on PBIS.

Horner, R. H., Kincaid, D., Sugai, G., Lewis, T., Eber, L., Barrett, S., . . . Johnson, N. (2014). Scaling up school-wide positive behavioral interventions and supports: Experiences of seven states with documented success. *Journal of Positive Behavior Interventions*, *16*, 197–208. doi:10.1177/1098300713503685

Horner, R. H., 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.

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Massar, M., McIntosh, K., & Mercer, S. H. (2017). *Factor analysis of an implementation fidelity measure for social behavior systems*. Manuscript submitted for publication.

McIntosh, K., Kim, J., Mercer, S. H., Strickland-Cohen, M. K., & Horner, R. H. (2015). Variables associated with enhanced sustainability of school-wide positive behavioral interventions and supports. *Assessment for Effective Intervention*, *40*, 184–191. doi:10.1177/1534508414556503

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

McIntosh, K., Mercer, S. H., Hume, A. E., Frank, J. L., Turri, M. G., & Mathews, S. (2013). Factors related to sustained implementation of schoolwide positive behavior support. *Exceptional Children*, *79*, 293–311.

McIntosh, K., Mercer, S. H., Nese, R. N. T., Strickland-Cohen, M. K., & Hoselton, R. (2016). Predictors of sustained implementation of school-wide positive behavioral interventions and supports. *Journal of Positive Behavior Interventions*, *18*, 209–218. doi:10.1177/1098300715599737

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066X.50.9.741

Nese, R., McIntosh, K., Nese, J., Hoselton, R., Bloom, J., Johnson, N., . . . Ghemraoui, A. (2016). Predicting abandonment of school-wide positive behavioral interventions and supports. *Behavioral Disorders*, *42*, 261–270. doi:10.17988/bd-15-95.1

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes: State-wide scale-up of school-wide positive behavior intervention and supports. *The Journal of Behavioral Health Services & Research*, *39*, 417–433. doi:10.1007/s11414-012-9290-2

Simonsen, B., Eber, L., Black, A. C., Sugai, G., Lewandowski, H., Sims, B., & Myers, D. (2012). Illinois statewide positive behavioral interventions and supports: Evolution and impact on student outcomes across years. *Journal of Positive Behavior Interventions*, *14*, 5–16. doi:10.1177/1098300711412601

Sugai, G., & Horner, R. H. (2009). Defining and describing school-wide positive behavior support. In W. Sailor, G. Dunlap, G. Sugai, & R. H. Horner (Eds.), *Handbook of positive behavior support* (pp. 307–326). New York, NY: Springer.

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

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

Turri, M. G., Mercer, S. H., McIntosh, K., Nese, R. N. T., Strickland-Cohen, M. K., & Hoselton, R. (2016). Examining barriers to sustained implementation of school-wide prevention practices. *Assessment for Effective Intervention*, *42*, 6–17. doi:10.1177/1534508416634624

Upreti, G., Liaupsin, C., & Koonce, D. (2010). Stakeholder utility: Perspectives on school-wide data for measurement, feedback, and evaluation. *Education and Treatment of Children*, *33*, 497–511. doi:10.1353/etc.2010.0001

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*, 21–29.