

Measuring School-wide Positive Behavior Support Implementation:

Development and Validation of the Benchmarks of Quality



Rachel Cohen
Don Kincaid
Karen Elfner Childs
University of South Florida

Abstract: School-wide positive behavior support (SWPBS) has been implemented in more than 4,000 schools as a means of addressing problem behavior in a systemic fashion. Preliminary outcomes (e.g., office discipline referrals, suspensions) indicate the effectiveness of SWPBS in decreasing school-wide behavior problems and creating a positive school climate. Although the results of a majority of the program evaluations yielded significant findings, there has been a lack of measurement of treatment fidelity, possibly due to the absence of expedient, effective assessment tools. This article describes the theoretical background and development, including a qualitative pilot study and psychometric properties, of the School-wide Benchmarks of Quality (BoQ; Kincaid, Childs, & George, 2005), a tool intended to measure the implementation of SWPBS. Descriptive data on the instrument, including internal consistency, test-retest reliability, interrater reliability, and concurrent validity, were collected and analyzed. Results indicate that the BoQ for SWPBS is a reliable, valid, efficient, and useful instrument for measuring the fidelity of implementation of the primary or universal level of PBS application in individual schools. Future considerations for evaluating the psychometric properties of the BoQ include extending the data collection and analysis to many more schools across multiple states.

School violence has been rated by the public as the top problem or concern in schools (Rose & Gallup, 2004; Mayer & Leone, 1999). A survey of a nationally representative sample of 1,000 teachers and 1,180 students in Grades 3 through 12 found that most teachers felt safe at schools but 11% had been the victims of violence on school property (Leitman & Binns, 1993). The same survey found that 50% of students felt safe but 23% reported being victims of violence and 22% were somewhat or very worried about being hurt at school. The presence of severe behavior problems in schools also is evidenced by a significant number of office discipline referrals (ODRs), suspensions, and expulsions for violating school rules, as well as staff member reports that indicate a desire to improve school discipline systems (Sugai et al., 1999; Taylor-Greene et al., 1997).

Many efforts to remediate system-wide discipline problems have been reactive in nature (University of Oregon, 2004). Such reactive or aversive strategies may result in an immediate reduction in problem behaviors, but such reductions are temporary, and problem behaviors often re-

occur (University of Oregon, 2004). In contrast, proactive approaches that emphasize teaching expectations and rewarding positive behavior are effective for the majority of students (Sugai et al., 1999). Furthermore, recent research has suggested that the best strategy for promoting responsible behavior in schools is to emphasize both the rules and the consequences of breaking the rules (Baer, Manning, & Shiomi, 2006). Findings of this nature have supported the need for a shift from a reactive to a proactive approach to discipline, such as school-wide positive behavior support (SWPBS).

School-wide PBS

SWPBS is an intervention intended to improve the climate of schools using system-wide positive behavioral interventions, including a positively stated purpose, clear expectations backed up by specific rules, and procedures for encouraging adherence to and discouraging violations of the expectations (Lewis & Sugai, 1999). SWPBS is a com-

ponent of a larger general initiative called *positive behavior support* (PBS), a “systems approach to enhancing the capacity of schools to adopt and sustain the use of effective practices for all students” (Lewis & Sugai, 1999, p. 4). Implementation of the PBS model involves three tiers of behavioral interventions—primary or universal, secondary, and tertiary—with the intensity of intervention intended to match the intensity of the problem behavior (Lewis & Sugai, 1999; Nelson, 2000; Sugai et al., 1999; Taylor-Greene et al., 1997; University of Oregon, 2004). This article will focus primarily on the universal level of SWPBS.

Evaluation of School-wide PBS

SWPBS is currently being practiced in more than 4,000 schools across the United States, and that number is expected to double in the next few years (U.S. Department of Education, 2005). With the rapid increase in the number of schools implementing SWPBS, there is an increasing need to develop an assessment of treatment integrity and outcomes (University of Oregon, 2004). Many studies have used discipline referral rates, suspension rates, and satisfaction reports to evaluate the overall effectiveness of SWPBS (i.e., student outcomes; Lewis & Newcomer, 2002; Taylor-Greene et al., 1997), and many schools have found an overall decrease in the number of discipline referrals 1 and 2 years after SWPBS implementation (Eber, Lewis-Palmer, & Pacchiano, 2001). Specifically, there have been significant decreases in disruption and fighting in the classroom and schoolyard (McCurdy, Mannella, & Eldridge, 2003) and in referrals for harassment (Metzler, Biglan, & Rusby, 2001). There also have been decreases in the number of days of out-of-school suspension (Scott, 2001) and the number of suspensions per day (Eber et al.).

Although most program evaluations yielded significant findings, some authors, such as Scott (2001), have noted that no measure of treatment fidelity was included in his study. Metzler et al. (2001) collected implementation data during Year 1 but not during subsequent years. Eber et al. (2001) used the number of teachers involved with the project as a measure of implementation but did not measure implementation of program components.

One reason for the lack of implementation measures in these studies may be that there is a paucity of implementation assessment instruments. The only widely used, validated tool for evaluating the implementation of SWPBS is the *School-Wide Evaluation Tool* (SET; Horner et al., 2004), a 28-item research-based observation and interview instrument. The SET contains seven subscales: Expectations Defined, Behavioral Expectations Taught, On-Going System for Rewarding Behavioral Expectations, System for Responding to Behavioral Violations, Monitoring and Decision-Making, Management, and District-Level Support.

Although the SET provides excellent information about implementation and has acceptable psychometric

properties, it is time intensive and requires on-site implementation. Our experience with the SET indicates that evaluators require 6 to 8 hrs of training (e.g., one to two school visits with an experienced evaluator) and that administration and scoring take 4 to 6 hrs per school (including travel time). SET also requires access to students, staff, and administrators. In addition, schools can score over 80% on the SET without having many of the critical features of SWPBS, such as lesson plans and an evaluation plan, in place. Finally, the information provided by the SET lends itself more to feedback about products (e.g., number of posters on walls) and knowledge of participants (e.g., percentage of students and staff who know the expectations) than to feedback about the implementation process that could help schools improve their programs.

With the dramatic increase in the number of schools implementing PBS, it is difficult to conduct on-site assessments at each school, and an instrument was needed that allowed school teams to assess their own strengths and weaknesses. The *School-wide Benchmarks of Quality* (BoQ; Kincaid, Childs, & George, 2005) self-report rating scale was developed to meet these two needs.

This article will describe the theoretical background, development, and psychometric properties of the BoQ. The BoQ was designed in the three stages described by McKenell (1974): qualitative pilot (development of instrument content), scale development pilot (development of the scale structure), and main survey (development of a context for the instrument within a conceptual network or the reliability and validity of the instrument). This article also will describe the methods and results used to examine the reliability (internal consistency, test-retest, interrater) and validity (concurrent) of the instrument. As Messick (1988) pointed out, instrument development and validation is an ongoing process. This article represents the framework and initial steps of an ongoing validation process for the BoQ.

Method

INSTRUMENT

Content

The BoQ is a 53-item rating scale that measures the degree of fidelity with which a school is implementing SWPBS. Based on Ponti, Zins, and Graden's (1988) argument that self-report measures are valid ways to assess the implementation of organizational interventions, this instrument was developed as a self-evaluation tool to allow school teams to review their progress toward implementing critical elements of PBS that are presented during training (Florida's Positive Behavior Support Project [FLPBS], 2004–2005). The critical elements correspond to the 10 subscales of the instrument: PBS Team, Faculty Commitment, Effective Discipline Procedures, Data Entry, Expectations and

Rules, Reward System, Lesson Plans, Implementation Plans, Crisis Plans, and Evaluation.

Development

The items were developed directly from the FLPBS training manual (FLPBS, 2004–2005), which is based on the critical elements of SWPBS (Lewis & Sugai, 1999). Items for the scoring guide were generated from the implementation goals documented in the training manual. After all the items were generated, approximately 20 key professionals (trainers and experts in PBS from several states) rated the importance of each item to the PBS process on a scale from 1 (*minimally important*) to 3 (*critically important*). These ratings were then used to establish the point values for each item.

Following item generation and rating, a technique called cognitive interviewing was used to find sources of response error in survey questions by asking survey respondents to think aloud while responding to the items (Schechter, Blair, & Vande Hey, 1996; Willis, 1999). This technique was used to ensure that respondents were interpreting the items as intended and to detect any problems that could occur with the instrument in a field setting. Three SWPBS coaches were selected from three Florida counties to participate in cognitive interviewing. They represented different races and genders: a Caucasian woman, an African American woman, and a Caucasian man. The interviewees were trained in the procedures and then asked to think aloud while completing the BoQs for their schools. The interviewer probed the coaches to clarify any unclear items or responses. The responses were summarized and used to revise the instrument.

The instrument was then piloted with 10 SWPBS coaches and teams in Florida. The BoQ was completed by each school coach or team, and feedback was provided on any items or directions that were unclear. These qualitative data were used to make additional revisions to the instrument, all of which were minor.

Administration and Scoring

The BoQ consists of three documents: the Coach Scoring Form, the Scoring Guide, and the Team Member Rating Form. The coach first completes the Coach Scoring Form using the Scoring Guide, which provides operational definitions of the scores for each item. Each team member then individually completes the Team Member Rating Form, a simplified version of the Coach Scoring Form that does not require the Scoring Guide. The raters instead indicate whether the content of each item is *not in place*, *needs improvement*, or is *in place*. After the coach and team members have completed the scoring forms, the coach compares his or her ratings with the team members' ratings, makes note of any discrepancies, and completes a Team Summary Report. Prior to or during the presentation of the Team Summary Report to the team, the coach

can discuss discrepancies and make any necessary adjustments to the score based upon additional information provided by team members.

The BoQ has a total possible score of 100. This score is derived from the three to eight items in each of the 10 subscales. Each item has a maximum value between 1 and 3 points, and points for the items are summed to obtain the total score.

SETTINGS AND PARTICIPANTS

Schools from two states that are implementing SWPBS at a statewide level participated in this investigation: 91 schools from Florida and 14 from Maryland. The BoQ was completed at all schools, while both the BoQ and the SET were completed at 34 of the 91 Florida schools and at 13 schools in Maryland.

The schools included represent a diverse sample. There were 44 elementary schools (Florida $n = 37$, Maryland $n = 7$), 35 middle schools (Florida $n = 30$, Maryland $n = 5$), 10 high schools (Florida $n = 9$, Maryland $n = 1$), and 16 center schools (Florida $n = 15$, Maryland $n = 1$). The percentage of non-White students in the schools ranged from 7% to 99% ($M = 45$, $SD = 21$) for Florida schools and 20% to 76% ($M = 45$, $SD = 21$) for Maryland schools. Free and reduced-price lunch status ranged from 17% to 97% for Florida schools ($M = 54$, $SD = 19$); lunch status data for Maryland were not available.

For the Florida Positive Behavior Support Project, completion of the BoQ was a required part of the end-of-year evaluation for the 2004–2005 school year. For the Maryland schools, the BoQ was included alongside the SET as part of the schools' regular year-end evaluation procedures. FLPBS provided the Maryland schools with a stipend for participation in this investigation.

PROCEDURE

The procedures for training and completing the assessments were slightly different for the two states.

Training

FLPBS staff trained Florida PBS coaches in the procedures for completing the BoQ in January 2005. The training included an explanation of the instrument and a practice session with a fictitious school. Coaches who did not attend the training session were provided with a CD containing the training presentation with a recorded voice explaining each slide.

FLPBS staff trained coaches in Maryland in the procedures for completing the BoQ in April 2005. This training was identical to that provided in Florida. A few coaches who did not attend the training received the training CD.

Data Collection

BoQ. The BoQ was completed during the end-of-the-year evaluation period. In Florida, this occurred between the end of March and the beginning of May. In Maryland, the evaluation period lasted from the end of April until the middle of June. For all 91 schools in Florida and all 14 schools in Maryland, the coach completed the BoQ following the previously described procedures.

SET. At 34 of the Florida schools, state PBS project staff completed a SET within 2 weeks of the BoQ. At 13 of the Maryland schools, the SET was completed by an outside evaluator within 2 weeks of the BoQ. The SET was not completed at one of the Maryland schools that completed the BoQ.

Reliability and Validity

Test-retest reliability. Fourteen coaches in each state, for a total of 28 coaches, completed the BoQ's Coach Scoring Form a second time within 1 week after he or she had completed it the first time. The coach was allowed to use the team members' rating forms to complete the BoQ both times, but only the responses from the Coach Scoring Forms were used in the analyses. To recruit coaches to complete the BoQ twice in Florida, a flyer was sent out to all coaches offering \$25 to participate in this component of the study.

Interrater reliability. In 21 Florida schools, individuals other than the coach who were familiar with the school (e.g., external coach or district coordinator) completed the BoQ within 2 weeks of the coach's completion of the BoQ. The second rater in Florida received a \$50 stipend for completing the CD-based training and the Scoring Guide. In 13 Maryland schools, individuals other than the coach completed the BoQ within 2 weeks. In all cases, the second rater used the team members' ratings and his or her own experience with the school to complete the instrument.

Concurrent validity. Concurrent validity, or the relationship between one instrument and another similar instrument, was measured by examining the BoQ's performance

in relation to the SET, which has demonstrated good psychometric properties. The administration of the SET was scheduled within approximately 2 weeks of the BoQ. The implementation of PBS is assumed to remain relatively stable over a 2-week period as most teams meet monthly and it is typically at the monthly meetings that the teams make changes to the implementation of PBS.

Results

DESCRIPTIVE STATISTICS

The means and standard deviations for all the instruments and administrations are presented in Table 1. The overall mean for the 105 schools that completed the BoQ was 69, with a standard deviation of 20. The scores ranged from 4 to 99. As each subscale had a different number of items and thus a different point total, the percentage of the total points obtained was calculated so that each subscale could be compared to one another. For example, if one school obtained a total of 8 points on a subtest with a total possible point value of 10, the score would be 80% for the subscale. The item, subscale, and total means and standard deviations are displayed in Table 2, along with the percentage of possible points obtained from each item and subscale.

RELIABILITY

Internal Consistency

Internal consistency was calculated using Cronbach's coefficient alpha, which was calculated for all BoQ subscales and the total score (see Table 3). The results document an overall alpha of 0.96, and the alphas for the subscales ranged from 0.43 to 0.87. With the exception of the first subscale ($\alpha = 0.43$), all alphas were above 0.70, which is the threshold set by Nunnally (1978) for making a determination that the items fit together on a scale. Pearson product-moment correlations also were used to examine all item-subscale and item-total correlations. The majority of these correlations fell between 0.40 and 0.70, indicating moderate correlation (see Table 3).

Table 1. Florida and Maryland Descriptive Statistics

Data source	BoQ score		Test-retest reliability		Interrater reliability		SET score	
	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>	<i>n</i>	<i>M (SD)</i>
Maryland	14	82.79 (9.53)	14	82.79 (10.53)	13	83.69 (9.51)	13	92.25 (7.74)
Florida	91	66.79 (19.65)	14	71.86 (13.10)	21	68.57 (17.26)	34	84.73 (10.15)
Both	105	69.33 (19.70)	28	77.32 (12.74)	34	74.35 (16.40)	47	86.81 (10.05)

Note. BoQ = Benchmarks of Quality (Kincaid et al., 2005); SET = School-wide Evaluation Tool (Horner et al., 2004).

Table 2. BoQ Item, Subscale, and Total Score Means, Standard Deviations, and Percentages of Points

Subscale	Item	Points	Total <i>M</i>	<i>SD</i>	%
1. PBS team	1	1	0.90	0.29	90
	2	3	2.56	0.68	85
	3	2	1.61	0.61	80
	4	1	0.89	0.32	89
	All	7	5.96	1.23	85
2. Faculty commitment	5	2	1.07	0.72	53
	6	2	1.21	0.57	60
	7	2	1.24	0.64	62
	All	6	3.51	1.59	59
3. Effective procedures for dealing with discipline	8	2	1.66	0.53	83
	9	1	0.94	0.23	94
	10	2	1.80	0.54	90
	11	3	2.55	0.76	85
	12	2	1.59	0.57	80
	13	1	0.88	0.33	88
	14	1	0.89	0.32	89
	All	12	10.30	2.39	86
4. Data entry and analysis established	15	3	2.61	0.74	87
	16	1	0.69	0.47	69
	17	1	0.85	0.36	85
	18	2	1.45	0.69	72
	19	2	1.09	0.72	54
	All	9	6.68	2.16	74
5. Expectations and rules developed	20	3	2.46	0.87	82
	21	3	2.46	0.81	82
	22	2	1.41	0.66	70
	23	1	0.81	0.39	81
	24	2	1.45	0.64	72
	All	11	8.58	2.46	78
6. Reward/recognition program established	25	3	2.10	0.96	70
	26	2	1.52	0.69	76
	27	3	2.30	0.89	77
	28	2	1.48	0.69	74
	29	1	0.89	0.32	89
	30	3	1.61	0.96	54
	31	1	0.46	0.50	46
	32	2	0.93	0.76	47
	All	17	11.29	4.36	66
	7. Lesson plans for teaching expectations/rules	33	2	1.14	0.7
34		1	0.70	0.4	70
35		2	0.99	0.79	50
36		2	0.88	0.73	44
37		1	0.47	0.50	47
38		1	0.32	0.47	32
All		9	4.50	2.99	50
8. Implementation plan		39	2	1.32	1.16
	40	2	1.05	0.67	52
	41	3	1.98	1.03	66
	42	2	0.94	0.72	47
	43	1	0.74	0.44	74
	44	2	0.89	0.67	44
	45	1	0.49	0.50	49
	All	13	7.41	3.62	57

(Table continues)

(Table 2 continued)

Subscale	Item	Points	Total <i>M</i>	<i>SD</i>	%
9. Crisis plan	46	1	0.88	0.33	88
	47	1	0.79	0.41	79
	48	1	0.90	0.31	90
	All	3	2.56	0.91	85
10. Evaluation	49	2	1.47	0.72	73
	50	2	1.50	0.65	75
	51	3	1.97	0.98	66
	52	3	1.78	1.07	59
	53	3	1.92	1.03	64
	All	13	8.70	3.52	67
Total BoQ	All	100	69.30	19.70	69

Note. *N* = 105 schools. BoQ = *Benchmarks of Quality* (Kincaid et al., 2005); PBS = positive behavior support.

The subscale–total and subscale–subscale correlations also were examined to determine the relationship between each subscale and between the subscales and the total score for the instrument (see Table 4). As with the SET, the BoQ subscales are intended to represent components of implementation. For the SET, Horner et al. (2004) suggested that the subscales should correlate with each other at least at the level of $r = 0.40$ but should not correlate much higher than $r = 0.80$ as they would then be measuring the same construct. For the BoQ, the subscale correlations range from 0.12 to 0.91 with an average of 0.51. A visual inspection of the correlations indicates that the majority of the scores fell within the ideal range, indicating an instrument with unique but related categories.

Test-Retest Reliability

For the 28 schools that completed the Coach Scoring Forms twice, Pearson product–moment correlations were calculated for the scores from Time 1 and Time 2, and the results indicated a high correlation of 0.94 ($p < 0.01$). Additionally, correlations were calculated for each of the subscales from Time 1 and Time 2. Results ranged from $r = 0.63$ to $r = 0.93$ (see Table 5).

To calculate test–retest reliability for the total score, a second method was used to obtain the percentage of agreement between the total score for Time 1 and the total score for Time 2. The lower score was divided by the higher score and multiplied by 100. For this sample, the average agreement was 97%.

Interrater Reliability

For the 34 schools for which two people completed the BoQ, Pearson product–moment correlations were calculated for the scores from both individuals, and the results indicated a high correlation of 0.87 ($p < .01$). The agree-

ment between raters was calculated by dividing the lower score by the higher score and multiplying by 100. For this sample, the average agreement was 89%.

CONCURRENT VALIDITY

To determine concurrent validity, the total scores on the BoQ were correlated with the total scores on the SET using Pearson product–moment correlations, and the results indicated a correlation of 0.51 ($p < 0.05$; see Table 6).

The BoQ scores averaged more than 15 points lower than the SET scores for the same schools in Florida and 9 points lower for schools in Maryland (see Table 1). A visual analysis of the data in a scatter plot (see Figure 1) indicates that only one school that scored above 80% on the BoQ failed to score above 80% on the SET. However, 13 schools that scored above 80% on the SET failed to score above 80% on the BoQ. Thus, the BoQ may be more able than the SET to discriminate among schools that are implementing with high fidelity because it covers critical features of SWPBS that are not covered by the SET.

Although the concurrent validity of the BoQ and the SET is important, the relationship of the BoQ scores to the standard measure of SWPBS outcome, change in office discipline referrals (ODRs), is even more important. For the 24 Florida schools with baseline measures, at least 2 years of intervention data, and a BoQ completed in 2005, it is clear that schools that had higher BoQ scores (70% or higher) tended to have greater decreases in the rate of ODRs than schools with lower BoQ scores (69% or lower). Seventy percent was used as a criterion because it was slightly higher than the Florida mean of 67%. In fact, from baseline through 2 years of implementation, schools with high BoQ scores had nearly three times the decrease in ODR rate than schools with low BoQ scores. Although the data are still preliminary, if this trend is maintained across mul-

Table 3. BoQ Item-Subscale, Item-Total, and Subscale-Total Internal Consistency Reliabilities

Subscale	Item	r_{ss}	r_{tot}	r_a
1. PBS team	1	0.25	0.20	0.43
	2	0.35	0.41	
	3	0.21	0.32	
	4	0.27	0.34	
2. Faculty commitment	5	0.48	0.56	0.75
	6	0.65	0.70	
	7	0.64	0.61	
3. Effective procedures for dealing with discipline	8	0.68	0.55	0.81
	9	0.61	0.44	
	10	0.58	0.37	
	11	0.62	0.44	
	12	0.64	0.60	
	13	0.48	0.53	
	14	0.52	0.59	
4. Data entry and analysis plan established	15	0.47	0.38	0.74
	16	0.34	0.42	
	17	0.51	0.29	
	18	0.72	0.56	
	19	0.58	0.59	
5. Expectations and rules developed	20	0.64	0.62	0.76
	21	0.60	0.63	
	22	0.58	0.55	
	23	0.38	0.45	
	24	0.47	0.58	
6. Reward/recognition program established	25	0.72	0.71	0.87
	26	0.80	0.76	
	27	0.73	0.71	
	28	0.70	0.65	
	29	0.56	0.58	
	30	0.71	0.75	
	31	0.31	0.32	
	32	0.62	0.66	
7. Lesson plans for teaching expectations/rules	33	0.75	0.56	0.87
	34	0.71	0.44	
	35	0.80	0.60	
	36	0.72	0.63	
	37	0.70	0.46	
	38	0.43	0.38	
8. Implementation plan	39	0.41	0.36	0.79
	40	0.65	0.57	
	41	0.60	0.67	
	42	0.66	0.69	
	43	0.48	0.58	
	44	0.60	0.61	
	45	0.47	0.52	
9. Crisis plan	46	0.72	0.41	0.83
	47	0.69	0.37	
	48	0.66	0.49	
10. Evaluation	49	0.28	0.37	0.83
	50	0.68	0.71	
	51	0.78	0.72	
	52	0.73	0.75	
	53	0.72	0.70	
TOTAL				0.96

Note. $N = 105$ schools. BoQ = *Benchmarks of Quality* (Kincaid et al., 2005); r_{ss} = BoQ item-subscale internal consistency reliability; r_{tot} = Item-total internal consistency reliability; r_a = Subscale-total internal consistency reliability; PBS = positive behavior support.

Table 4. BoQ Subscale Correlation Matrix

Subscale and total BoQ	Total	1	2	3	4	5	6	7	8	9	10
Total	1.00	0.55	0.77	0.72	0.68	0.83	0.91	0.69	0.83	0.50	0.88
1. PBS team	—	1.00	0.42	0.64	0.46	0.38	0.43	0.37	0.39	0.12	0.33
2. Faculty commitment	—	—	1.00	0.50	0.54	0.65	0.63	0.53	0.65	0.33	0.63
3. Effective procedures for dealing with discipline	—	—	—	1.00	0.55	0.55	0.62	0.39	0.42	0.37	0.51
4. Data entry and analysis plan established	—	—	—	—	1.00	0.47	0.54	0.33	0.48	0.25	0.59
5. Expectations and rules developed	—	—	—	—	—	1.00	0.77	0.44	0.63	0.49	0.73
6. Reward/recognition program established	—	—	—	—	—	—	1.00	0.55	0.73	0.44	0.83
7. Lesson plans for teaching expectations/rules	—	—	—	—	—	—	—	1.00	0.65	0.21	0.47
8. Implementation plan	—	—	—	—	—	—	—	—	1.00	0.36	0.73
9. Crisis plan	—	—	—	—	—	—	—	—	—	1.00	0.51
10. Evaluation	—	—	—	—	—	—	—	—	—	—	1.00

Note. $N = 105$ schools. BoQ = *Benchmarks of Quality* (Kincaid et al., 2005).

Table 5. Test-Retest and Interrater Reliability Correlations by Subscale

Subscale	Test-retest reliability r	Interrater reliability r
1	0.90	0.68
2	0.79	0.58
3	0.76	0.65
4	0.78	0.78
5	0.90	0.67
6	0.92	0.89
7	0.90	0.72
8	0.63	0.52
9	0.87	0.79
10	0.93	0.79

multiple schools in multiple states, it would certainly support the practical usefulness of the BoQ.

Discussion

UTILITY OF THE BOQ

The results of our evaluation indicate that the *School-wide Benchmarks of Quality* for SWPBS is a reliable, valid, efficient, and useful instrument for measuring the degree of

implementation of the primary or universal level of PBS application within individual schools. The high test-retest reliability (above 90%) indicates that the BoQ is a stable instrument, and the high interrater reliability (also above 90%) indicates that the BoQ process, including the Scoring Guide, allows for accurate and consistent scoring across different evaluators.

The BoQ's moderate correlation with the SET was not unforeseen. Because the process of primary or universal implementation of PBS shares many common features across the country (an effective response to discipline incidents, school expectations and rules, teaching students, etc.), it is not surprising that the BoQ measures many of the same features of SWPBS as the SET. However, the BoQ measures many of those areas with more specificity than the SET, and it measures other areas related to the implementation that are not covered by the SET. Although seven of the subscales on the SET and BoQ represent similar elements of PBS, some subscales are included on one tool but not the other. For instance, the SET includes a subscale on district support that is not included on the BoQ. The two questions on this subscale are about PBS funding in the school budget and the identification of a district PBS liaison. Twenty-eight out of the 34 schools in the Florida PBS project that were administered the SET received a score of 100% on this section; therefore, this section may contribute to higher scores on the SET than the BoQ. The BoQ, for its part, has four sections that are not included in the SET: faculty buy-in, lesson plans, crisis plans, and evaluation. Although the SET includes a section on teaching the expectations, the section on the BoQ focuses on the quality of lesson plans. On average, Florida schools re-

Table 6. Test-Retest, Interrater Reliability, and Concurrent Validity Correlations by State

Data source	Test-retest reliability		Interrater reliability		Concurrent validity ^a	
	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>	<i>n</i>	<i>r</i>
Florida	14	0.98	19	0.86	29	0.45
Maryland	14	0.83	13	0.76	13	0.49
Both	28	0.94	32	0.87	42	0.51

^aIndicates correlations with *School-wide Evaluation Tool* (Horner et al., 2004).

ceived the lowest percentage of points on this BoQ subscale ($n = 91$, $M = 46\%$, $SD = 33$). As these items were not reflected in the SET and they tend to be scored lower, they likely influence the lower scores on the BoQ. Further research may be necessary to identify the degree to which these areas contribute to high or low fidelity of implementation of the PBS process.

The preliminary data regarding the association between BoQ scores and reductions in ODRs also is encouraging. If reductions in ODRs are related to the fidelity of implementation, then the correlation between higher BoQs and larger decreases in ODRs should continue. Data gathered from many more schools over the next few years should provide further evidence of this association.

IMPLICATIONS FOR PRACTICE

The BoQ has several ease-of-use advantages over the SET. First, scorers can learn to use the BoQ instrument accurately with little training. Training may take as little as 30 min and can be done in person, via CD, or on the web. The well-organized protocol for each step in the BoQ process and the precise scoring criteria for each item are helpful in simplifying the assessment process. Second, the BoQ may require as little as 10 min from team members and 60 to 90 min from the coach for completion. SETs may require 3 to 6 hrs of an evaluator's time (travel, scoring, on-site time) and access to team members, students, and administrators. Third, the consistency of results across the two states (Florida and Maryland) indicates that the areas measured by the BoQ are not unique to a training or implementation approach used in one state. More importantly, the BoQ remained a reliable and valid tool regardless of the type of respondent using the instrument. In Maryland, BoQ respondents were school personnel trained in implementing the SET in multiple schools. In Florida, most of the respondents were coaches who had never used or been trained in the SET. Finally, the BoQ holds promise as an instrument that can assist states that are rapidly expanding their implementation efforts from a few schools to hundreds of schools. The cost of training and paying for SET

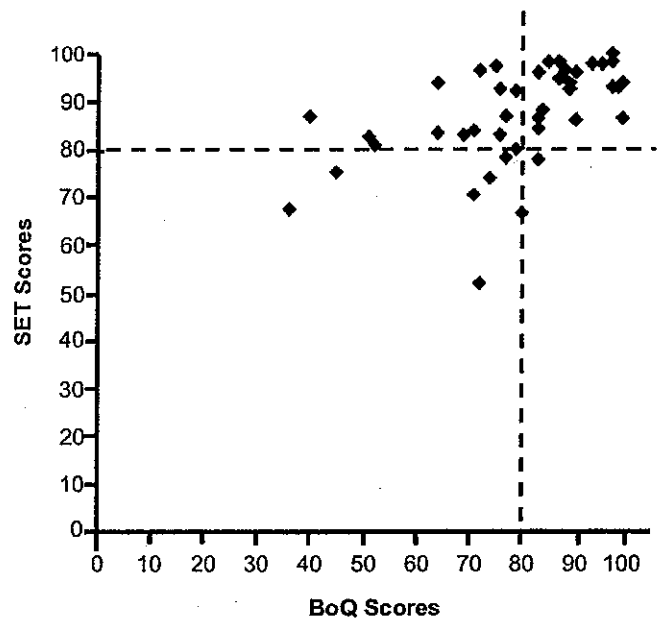


Figure 1. Scatter plot of SET and BoQ scores. *Note.* SET = *School-Wide Evaluation Tool* (Horner et al., 2004); BoQ = *Benchmarks of Quality* (Kincaid et al., 2005).

evaluators may be prohibitive for many state projects. The relative ease of training and the local team-level use of the BoQ should contain evaluation costs and provide reliable and valid data for state-level evaluation and data-based decision making.

The scores generated by the BoQ on average were lower than the SET scores for the same Florida and Maryland schools, resulting in a wider range of scores for those schools with a higher implementation. Therefore, obtaining a wider range of scores, particularly at the top end of the spectrum (scores over 70), allows evaluators to more clearly discriminate relative fidelity of implementation. Our experience with the SET indicates that it may be possible for schools that score above 80 on the SET to be missing some key ingredients that are essential for sustained implementation. The BoQ may provide a finer analysis of

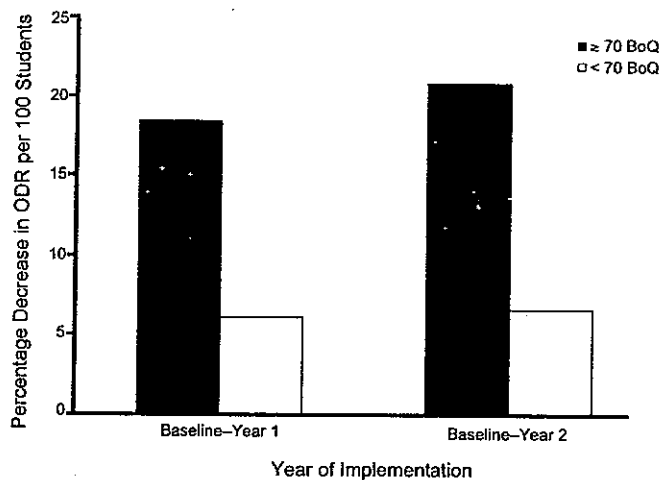


Figure 2. Percentage decrease in office discipline referrals (ODRs) per 100 students for schools scoring ≥ 70 and those scoring < 70 on the *Benchmarks of Quality* (BoQ; Kincaid et al., 2005).

critical areas necessary for successful implementation of SWPBS.

The BoQ was also developed to provide coaches with very clear information about areas of strength and weakness in their implementation efforts. The Team Summary Report provides information to help the team celebrate their successes and plan steps to address deficits in their implementation efforts. Coaches indicated that this was one of the most useful aspects of the BoQ.

When making decisions about whether to use an observation-interview tool like the SET or a self-report tool like the BoQ, it is important to consider such factors as the amount of time available, the number of schools being evaluated, and the resources available. If the evaluator has the time and resources, both tools can be used to cross-validate each other's results. The SET can provide valuable objective information that can be used to validate the reporting on the BoQ, and the BoQ can be used as a self-study tool for teams to determine additional areas in need of improvement that were not identified on the SET. Both implementation tools are an integral component of state PBS projects' and SWPBS trainers' data-based decision-making and evaluation processes.

LIMITATIONS

As mentioned previously, a major limitation of the current study for determining the psychometric properties of the BoQ is the small sample size. Although the preliminary data are compelling, further analysis of the BoQ instrument will require a larger sample size across multiple sites and times. At a minimum, the analysis of several hundred

SETs and BoQs will be necessary for the next step in the validation process.

Another limitation of the BoQ is the potential for rater bias. Since the BoQ is a self-report tool, a particular coach may not accurately gauge the performance of the team or school, perhaps because of limited exposure to the daily activities of the team. The BoQ was designed to minimize this source of bias by requiring the coach to view ratings of the team members and discuss discrepancies with the team prior to submitting a final score. In addition, the scoring rubric provides the coach with assistance in assigning scores for individual items. It is hoped that these two strategies will control for most of the respondent bias.

Another limitation of the BoQ is a lack of on-site observation, one of the strongest features of the SET. Future evaluation of the BoQ may include the use of on-site staff and student interviews (such as those included in the SET) as a reliability check for a sample of BoQs. In addition, it may be useful to evaluate on-site interviews with the coach or a district coordinator as a supplement to the BoQ. Such a process may make the BoQ an even better tool for evaluating the fidelity of implementation of SWPBS.

CONCLUSION

This initial investigation into the quantitative soundness and usefulness of the *School-wide Benchmarks of Quality* for SWPBS found it to be a reliable and valid tool for assessing the implementation of universal or primary PBS in a school. Although the preliminary data are encouraging, identifying the BoQ's psychometric properties will require an expanded sample size. In addition, the literature points to the need to develop other measurement tools with established psychometric properties for assessing the PBS process at the primary, secondary, and tertiary levels. Other evaluation tools may be necessary for assessing specific aspects of implementation such as administrative support, team buy-in, team process, and coaching effectiveness. Although such components appear to be directly related to effective implementation, we have few measures to assess these components and even fewer with established psychometrics. Finally, the field of PBS is encouraged to consider the relationship between evaluation tools and direct measures of change such as ODRs, suspensions, and academic performance. Evaluation tools with both established psychometric properties and practical application will assist the field of PBS in modeling data-based decision making at the local, district, state, and federal levels.

ABOUT THE AUTHORS

Rachel Cohen, PhD, completed her doctoral dissertation on the factors that influence the implementation of school-wide

positive behavior support. She currently works as an intervention specialist in a northern suburb of Chicago. Don Kincaid, EdD, is a research professor at the University of South Florida and principal investigator on several positive behavior support projects. Karen Elfner Childs, MA, is currently the research and evaluation coordinator for Florida's Positive Behavior Support Project. Address: Rachel Cohen, 2113 W. Addison Ave. #1, Chicago, IL 60618; e-mail: rachelmcohen@gmail.com

REFERENCES

- Baer, G. G., Manning, M. A., & Shiomi, K. (2006). Children's reasoning about aggressions: Differences between Japan and the United States and implications for school discipline. *School Psychology Review, 35*, 62-77.
- Eber, L., Lewis-Palmer, T., & Pacchiano, D. (2001, February). School-wide positive behavior systems: Improving school environments for all students including those with EBD. Paper presented at the 14th Annual Research Conference, Tampa, FL.
- Florida's Positive Behavior Support Project. (2004-2005). *Team training on school-wide positive behavior support*. Unpublished training manual, University of South Florida.
- Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Bolland, 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., & George, H. (2005). *School-wide benchmarks of quality*. Unpublished instrument, University of South Florida.
- Leitman, R., & Binns, K. (1993). *The American teacher 1993: Violence in America's schools: Metropolitan Life survey*. New York: Louis Harris. (ERIC Document Reproduction Service No. ED397190)
- Lewis, T. J., & Newcomer, L. L. (2002). Examining the efficacy of school-based consultation: Recommendations for improving outcomes. *Child & Family Behavior Therapy, 24*, 165-181.
- Lewis, T. J., & Sugai, G. (1999). Effective behavior support: A systems approach to proactive school-wide management. *Focus on Exceptional Children, 31*, 1-17.
- Mayer, M. J., & Leone, P. E. (1999). A structural analysis of school violence and disruption: Implications for creating safer schools. *Education & Treatment of Children, 22*, 333-356.
- McCurdy, B. L., Mannella, M. C., & Eldridge, N. (2003). Positive behavior support in urban schools: Can we prevent the escalation of antisocial behavior? *Journal of Positive Behavior Interventions, 5*, 158-179.
- McKenna, A. C. (1974). *Surveying attitude structures*. Amsterdam: Elsevier.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer (Ed.), *Test validity* (pp. 33-48). Hillsdale, NJ: Erlbaum.
- Metzler, C. W., Biglan, A., & Rusby, J. C. (2001). Evaluation of a comprehensive behavior management program to improve school-wide positive behavior support. *Education & Treatment of Children, 24*, 448-479.
- Nelson, J. R. (2000). Educating students with emotional and behavioral disabilities in the 21st century: Looking through windows, opening doors. *Journal of Education and Treatment of Children, 23*, 204-220.
- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Ponti, C. R., Zins, J. E., & Graden, J. L. (1988). Implementing a consultation-based service delivery system to decrease referrals for special education: A case study of organizational considerations. *School Psychology Review, 17*, 89-100.
- Rose, L. C., & Gallup, A. M. (2004). The 36th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 86*. Retrieved June 2005 from <http://www.pdkintl.org/kappan/k0409pol.htm>
- Scott, T. M. (2001). A school-wide example of positive behavior support. *Journal of Positive Behavior Interventions, 3*, 88-94.
- Schechter, S., Blair, J., & Vande Hey, J. (1996). Conducting cognitive interviewing to test self-administered and telephone surveys: Which methods should we use? In *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 10-17). Alexandria, VA: American Statistical Association.
- Sugai, G., Horner, R. H., Dunlap, G., Hieneman, M., Lewis, T. J., Nelson, C. M., et al. (1999). *Applying positive behavioral support and functional behavioral assessment in schools* (Technical Assistance Guide 1, Version 1.4.4). Eugene: University of Oregon, Center on Positive Behavioral Interventions and Support.
- Taylor-Greene, S. J., Brown, D., Nelson, L., Longton, J., Gassman, T., Cohen, J., et al. (1997). School-wide behavioral support: Starting the year off right. *Journal of Behavioral Education, 7*, 99-112.
- University of Oregon, Center on Positive Behavioral Interventions and Supports. (2004). *School-wide positive behavior support: Implementers' blueprint and self-assessment*. Eugene, OR: Author.
- U.S. Department of Education, Office of Special Education Programs. (2005). *Technical assistance center on positive behavioral interventions and supports: Final report*. Washington, DC: Author.
- Willis, G. B. (1999, August). *Cognitive interviewing: A "how to" guide*. Paper presented at the meeting of the American Statistical Association, Baltimore, MD.

Action Editor: Joshua K. Harrower

