Hammill Institute on Disabilities

Journal of Positive Behavior Interventions
2023, Vol. 25(2) 83–94

DOI: 10.1177/10983007211071127
jpbi.sagepub.com
SAGE

# Development and Validation of a Tool to Examine Program-Wide Implementation of the Pyramid Model

**Christopher Vatland, PhD[1], Erin E. Barton, PhD[2], Lam Pham, PhD[3], Lise Fox, PhD[1], Mary Louise Hemmeter, PhD[2], and Gary Henry, PhD[4]**

**Abstract**

In recent years, there has been increased attention regarding systems-level implementation to support the sustained use of evidence-based interventions and supports in authentic early childhood settings. With this comes a need to accurately measure implementation fidelity of the critical features within a framework as well as individual practices. Program-Wide Support for Pyramid Model Implementation (PWS-PMI) provides an approach for early childhood programs to develop such a framework that can underpin evidence-based practices in their classrooms. This article describes an evaluation of the technical properties of the Supporting Program-wide Implementation Fidelity Instrument (SPIFI), a fidelity tool that was developed to be used by typical evaluators to measure PWS-PMI in these settings. Findings suggest that the instrument reliably demonstrated construct validity when used by typical evaluators to assess PWS-PMI and provides initial validation of the SPIFI as an objective measure for use in evaluative research and technical assistance.

**Keywords**

school-wide, intervention(s), change, systems, program, assessment, early childhood

Effectively measuring and interpreting implementation fidelity is critical for replicating quality practices (Fixsen et al., 2005; Wasik et al., 2013). With the high numbers of expulsions, suspensions, and exclusionary practices that disproportionately impact children of color and with disabilities in early childhood settings (Meek & Gilliam, 2016), both educators and policymakers increasingly recognize the need for ensuring implementation fidelity of effective practices to support young children's social-emotional growth and address challenging behavior (U.S. Department of Health and Human Services and U.S. Department of Education, 2014).

The Pyramid Model (Fox et al., 2003; Snyder et al., 2022) was developed to guide early educators on evidence-based practices that promote social-emotional development and address challenging behavior in young children. The Pyramid Model is a multitiered framework built on evidence-based practices related to teaching social-emotional skills (Domitrovich et al., 2012), implementing positive behavior supports (Blair et al., 2010), and providing instruction in early childhood settings (Burchinal et al., 2010). The Pyramid Model has been examined in two randomized controlled trials (Hemmeter et al., 2016, 2021) in which preschool teachers received training and coaching to implement Pyramid Model practices. In both studies, implementation was associated with improved social skills for all children, and improved social skills and decreased challenging behavior in children identified as having social, emotional, or behavioral concerns at the onset of the study. The variation in the teachers' implementation of the model and the positive correlation between classroom implementation and improved outcomes led researchers to conclude that programs should develop systems to facilitate implementation using a model of program-wide support (Fox & Hemmeter, 2009; Hemmeter et al., 2013). Through program-wide support for the Pyramid Model, programs develop systems to reduce variation and increase fidelity of implementation at the program, classroom, and individual levels (Fox & Hemmeter, 2009). Similar to systemic efforts in supporting Positive Behavior Interventions and Supports (PBIS) in

[1]University of South Florida, Tampa, USA
[2]Vanderbilt University, Nashville, TN, USA
[3]North Carolina State University, Raleigh, USA
[4]University of Delaware, Newark, USA

**Corresponding Author:**
Christopher Vatland, Department of Child and Family Studies, College of Behavioral and Community Sciences, University of South Florida, MHC 2105, 13301 Bruce B Downs Blvd., Tampa, FL 33613, USA.
Email: cvatland@usf.edu

**Action Editor:** Grace Gengoux

school-age populations, program-wide efforts are designed to sustain and scale-up implementation of the framework (Sugai & Horner, 2002).

Implementation science recognizes that core systemic components (i.e., implementation drivers) contribute to both building and sustaining systemic support for desired practices in authentic settings (Fixsen et al., 2009; Metz et al., 2015). In early childhood systems, there are unique features that can and often do affect sustained implementation of the Pyramid model, including (a) staff capacity and support, (b) administrative competency in approaches to systemic implementation, (c) the developmental needs of preschool-age children, and (d) financial constraints (Fox & Hemmeter, 2009). A program-wide approach to implementing the Pyramid Model encompasses several key implementation drivers, including (a) a teaming structure and activities that involve professionals (e.g., therapists, behavior specialists); (b) active engagement with families around program-wide practices and individual child supports; (c) ongoing assessment and use of data to examine behavioral policies; (d) staff support, guidance, and buy-in related to implementation; and (e) ongoing professional development that includes classroom coaching (Quesenberry et al., 2011). The leadership team is composed of administrators, teachers, other professionals (e.g., therapist, coach, behavior specialist), and family members (Fox & Hemmeter, 2009) to guide the program using ongoing data related to both implementation and outcomes. These data inform decisions related to staff support and professional development. Although there are several valid fidelity measures of early childhood classroom and program practices, there are no validated, objective, and comprehensive measures of the overarching, program-wide systemic drivers that facilitate these practices.

### *Systemic Measurement for Pyramid Model*

Two tools have been developed to assess implementation of the Pyramid Model in early childhood settings. The Early Childhood Program-wide PBS Benchmarks of Quality (ECBoQ; Fox et al., 2017) provides a team tool to self-assess Pyramid Model implementation. The ECBoQ, however, does not provide an objective and reliable measure of systems implementation fidelity. The Preschool-wide Evaluation Tool (PreSET) evaluates critical universal features of PBIS implementation in early childhood programs (Steed et al., 2012); however, it does not reflect some of the most recent conceptualization of systems change and mechanisms that enable and sustain these changes (e.g., teaming structures and procedures; bidirectional engagement with staff and families; and multilevel data collection, analysis, and response) or mechanisms that facilitate more targeted and individualized intervention.

The Supporting Program-wide Implementation Fidelity Instrument (SPIFI) was developed as an objective measure to assess implementation fidelity related to critical features for building and sustaining the Pyramid Model in early childhood programs. The SPIFI focuses on program-wide implementation for a systemic framework that researchers (Durlak, 2016) have highlighted as critical for scale-up. The SPIFI provides a comprehensive assessment of the features needed for full implementation of all tiers (i.e., universal, targeted, and tertiary levels) of the Pyramid Model. In addition to procedures related to quality behavior support and intervention, the SPIFI includes indices related to leadership team practices, staff buy-in, strategies to support and promote family involvement, support for program-wide expectations, systems for designing and delivering effective interventions to children with social-emotional concerns and persistent challenging behavior, professional development and staff support, and data-based decision-making that includes implementation and behavior measures.

## Method

The development and evaluation of the SPIFI occurred in three phases: (a) developing indicators, (b) obtaining feedback from experts, and (c) gathering data from actual programs.

### *SPIFI Development*

The SPIFI was developed as part of a U.S. Department of Education grant-funded effort by the first four authors in consultation with five national experts in the areas of early childhood education, positive behavior support, and program-wide implementation of the Pyramid Model. The initial instrument reflected program features that were identified as necessary and beneficial for sustained implementation of Pyramid Model practices. In the initial draft, there were 77 total items identified that represented the indicators of implementation across nine domains: (a) leadership team composition (e.g., roles of members), (b) leadership team activities (e.g., frequency of meeting, action planning), (c) staff buy-in (e.g., orientation, ongoing data related to buy-in), (d) development and implementation of program-wide expectations (e.g., process for development, teaching), (e) procedures for developing behavior support plans, (f) staff support plan (e.g., professional development, coaching), (g) family engagement around the program-wide plan (e.g., input on expectations, ongoing bidirectional communication), (h) family engagement related to supports for individual children (e.g., involvement in planning), and (i) data-based decision-making (e.g., implementation data, outcome data). Once the domains and indicators were fully developed, the authors developed methods (e.g., interview,

observation, record review) for extracting data to reliably score the tool. Fifteen sources of evidence were present in this draft version: three interviews, two observations, and 10 permanent products. The team sought input from the national expert panel on the scope of the indicators and how accurately they represent program-wide implementation, sources of evidence, scoring criteria, balance of weighting, and ability to differentiate between levels of implementation using the instrument. The expert panelists suggested additional indicators, reframing of indicators within each domain, and guidance on scoring indicators and additional sources of evidence. These suggestions were integrated into a revised draft.

The researchers shared the refined draft tool with five experts who had extensive experience coaching early childhood program staff to implement Pyramid Model and providing ongoing monitoring of program performance. Cognitive interviewing was used with the experts to assess the understandability and usability of the SPIFI and to ensure minimal errors in interpreting individual indicators (Willis, 2004). These interviews were conducted using the "think aloud" technique, in which the interviewee was asked to vocalize their thoughts while responding to each indicator. When all interviews were completed, results were reviewed with the research team to determine necessary changes to the SPIFI and related documents.

The resulting SPIFI utilizes a multimodal evaluation process that includes three sources of evidence: interviews, observations, and permanent product review. To increase the reliability of the ratings, the procedures are documented in a 66-page manual that also includes a scoring guide, scoring sheets, permanent product rating forms, and interview and observation forms. The instrument has 82 indicators related to the aforementioned nine domains (see Table 1). The increased number of indicators reflected a recognized need to separate some of the earlier indicators that looked at the presence of multiple mechanisms and to further clarify specific activities that were attributed to implementation within each domain. The indicators are arranged in a hierarchical fashion within each domain with seven to 12 indicators per domain. The guide describes how to conduct interviews, observations, and product reviews and how to use the information to score indicators.

## Psychometric Integrity Study of the SPIFI

To assess the generalizability, robustness, and range of applicability of the SPIFI protocols and procedures as specified in the data collection manual, we collected data in a diverse group of early childhood programs. The psychometric integrity study of the SPIFI used data from 16 early childhood programs that participated in a randomized control trial in two large metropolitan areas in the southeast United States. These programs were participating in a larger pilot study examining the effectiveness of an intervention to implement systems that facilitate program-wide support for Pyramid Model implementation (PWS-PMI; Hemmeter et al., 2022). Project staff collected SPIFI data from participating control and intervention early childhood programs 3 times over the course of 1 year. The assignment of the programs to treatment or control was anonymized to the project staff collecting the data.

*Evaluator training.* All evaluators received a draft version of the protocol 2 weeks prior to training. They then participated in a 2-hr training on procedures for data collection and scoring with a draft version of the tool and scoring guide. Following training, evaluators had an opportunity to practice data collection at a local site and then debrief and discuss the methods and procedures. Following the initial training and practice session, evaluators conducted live observations until 80% interrater agreement was reached.

### *Primary measure*

*SPIFI.* SPIFI data collection involves between 2 and 4 hr of on-site time by the evaluator who visited each site before the school day began and then completed all on-site evaluation that day or in a subsequent visit less than 1 week later. The evaluation began with the administrator interview. Other interviews (with teachers, coaches, behavior support staff) occurred when it was convenient for the interviewees. The program observation occurred at a time when there were children in common areas and classroom observations were scheduled at convenient times for teachers. Adding to interviews and observations, evaluators reviewed the following permanent products when available (e.g., behavior support plans, coaching notes, family engagement products, staff poll, training log, team meeting minutes). Data across these sources were summarized in the Document Review section of the SPIFI.

The SPIFI Scoring Indicators and Clarifications provided guidance for how to score each of the indicators as present or absent. Each domain included indicators in four columns, arranged from insufficient implementation to full implementation. Once indicators were scored, evaluators were able to tabulate the domain score on a scale of 1 to 7. A score of 1 indicates "Insufficient Implementation," 3 indicates "Emerging Implementation," 5 indicates "Partial Implementation," and 7 indicates "Full Implementation." If indicators that were lower in the hierarchy were marked as absent, then rating would be attributed to the highest level of implementation that was marked as present. If all items were present in a lower level and the majority, but not all indicators, in a higher level were present, then the domain would receive the intermediate score (e.g., all indicators in Level 3 and half in Level 5 would warrant a score of 4).

**Table 1.** Average Scores for Overall SPIFI and for Each Indicator Across Domains

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Measure & domain | Wave 1 | Wave 2 | *p* value (Wave 1 vs. Wave 2 of pilot sample) | Both waves |
| Overall SPIFI scores | | | | |
| Sum of all items (Possible range: 9–63) | 12.3 (3.13) | 19.7 (10.4) | .0087 | 15.8 (8.23) |
| Sum of All Indicators (Possible range: 0–82) | 12.4 (8.82) | 28.3 (24.4) | .018 | 19.9 (19.3) |
| Scores per domain (Possible range: 1–7) | | | | |
| Domain 1: Leadership team | 1.06 (0.24) | 2.10 (1.96) | .037 | 1.55 (1.43) |
| Domain 2: Leadership team activities | 1 (0) | 2.20 (1.77) | .0088 | 1.56 (1.34) |
| Domain 3: Staff buy-in | 1 (0) | 2 (1.25) | .0025 | 1.47 (0.98) |
| Domain 4: Program-wide expectations | 1.24 (0.66) | 2.40 (1.47) | .0060 | 1.78 (1.24) |
| Domain 5: Behavior support plan | 1.56 (0.98) | 2.10 (1.38) | .21 | 1.81 (1.20) |
| Domain 6: Staff support plan | 1 (0) | 2.10 (1.37) | .0023 | 1.52 (1.07) |
| Domain 7: Family engagement | 1 (0) | 2.07 (1.82) | .022 | 1.50 (1.34) |
| Domain 8: Family engagement in individual students | 3.41 (2.00) | 3.40 (2.27) | .99 | 3.41 (2.10) |
| Domain 9: Data-based decision-making | 1.03 (0.12) | 1.33 (0.52) | .027 | 1.17 (0.39) |
| Observations | 17 | 15 | | 32 |

*Note.* Standard deviations in parentheses. *p* values in the third column are from *t* tests comparing Waves 1 and 2 of the pilot study sample.
SPIFI = Supporting Program-wide Implementation Fidelity Instrument.

*Measures used for validity assessment*

*Teaching pyramid observation tool (TPOT).* The TPOT (Fox et al., 2014) measures use of Pyramid Model practices by teachers. The TPOT is sensitive to changes in teachers' practices related to the Pyramid Model (Snyder et al., 2013). A generalizability theory study (G-study; Shavelson & Webb, 1991) with 50 preschool classrooms showed minimal error variance (5%) attributed to occasions and raters and a .97 generalizability coefficient.

*Classroom assessment scoring system (CLASS).* The CLASS (Pianta et al., 2008) instrument examines the quality of teacher–child interactions in the classroom. Composite domain scores on the CLASS range from 1 to 7 and interrater score reliability range from 78.8 to 96.9. Internal consistency score reliability estimates range from .79 to .91 in preschool classrooms. Confirmatory factor analyses, using data from five samples, support the theoretical structure of the measure. Structure coefficients ranged from .69 to .96. Goodness-of-fit indices ranged from .89 to .97 across samples and comparative fit indices ranged from .93 to .96.

*PreSET.* The PreSET (Steed et al., 2012) evaluates the fidelity of the universal PBIS practices in early childhood settings. The range of scale scores for PreSET is 0 to 100 and scores indicate the percentage of indicators observed within each subscale. These scale scores are standardized with a mean of 0 and a standard deviation of 1. Interrater reliability (IRR) yielded an agreement of 95% and an overall *k* value of .80 with good internal consistency ($\alpha = .91$) and strong indicator-subscale correlations (mean of .56 and a median of .58; Steed & Webb, 2013).

*Demographics.* We used a questionnaire to ask program leaders to document the number of (a) children per classroom, (b) number with an Individualized Education Program (IEP) or Individualized Family Service Plan (IFSP), (c) number who were multilanguage learners, (d) number with tuition fees, (e) number in Head Start/EHS, and (f)

number by gender and race. The questionnaire also asked teachers their race, ethnicity, background, years of experience, and social emotional curricula used.

*Data collection procedures.* SPIFI data were conducted with pilot study programs prior to leadership team training (Wave 1) and again 6 months after an external coach had worked with the program (Wave 2). For a randomly selected subsample of pilot site visits (25%), a second observer participated in the SPIFI data collection and independently scored the SPIFI.

Evaluators collected the SPIFI data as prescribed above. Evaluators began with the director interview, which provided information about program operations and expectations that informed other interviews and observations. Classroom observations and program observation occurred while children and adults were engaged in group and individual activities. Interviews were scheduled when teachers and staff were available across the half-day. All SPIFI data collection was conducted on-site except for the family interviews, which were conducted by phone or in-person fewer than 2 weeks following the on-site data collection. Programs with complete SPIFI and demographic data were included. The TPOT, CLASS, and PreSET data used in this study were collected for all sites in the pilot study in the same month as the SPIFI.

*Data analysis procedures.* We incorporated descriptions of statistical analysis methods into the "Results" section to better connect the analytic decisions with findings. As the SPIFI is a program-level fidelity of implementation instrument (Fox et al., 2003; Hemmeter et al., 2016), each program serves as a unit of primary analytic interest. We used the SPIFI data from the pilot study sample to examine IRR and conduct exploratory factor analyses (EFAs) of the underlying latent constructs measured by the SPIFI. Then, using classroom-level data, we assessed convergent validity using hierarchical linear models to estimate the association between SPIFI scores and other program-wide implementation measures (e.g., PreSET) and classroom measures of related constructs (e.g., CLASS, TPOT). Factor analyses and multilevel models used to assess convergent validity were estimated using Stata 15 (StataCorp, 2017). IRR measures were estimated using the *irr* package in *R* (Gamer et al., 2010).

## Results

### Descriptive Program Characteristics

Complete descriptions of program and teacher characteristics in the pilot study are available as a supplemental file. The pilot study programs served mainly Black (44%) and White (33%) children. About 14% of children are dual language learners, 73% of programs charge a tuition fee, and 33% of the sample are Head Start programs. Teachers were all female (100%) and the majority identify as Black (59%). On average, teachers had been working in their current position for approximately 5 years.

### Descriptive SPIFI Scores

Table 1 provides means and standard deviations of SPIFI scores from 16 programs. Reflecting typical ways in which the SPIFI might be scored, Table 1 presents the overall SPIFI score as either the sum of the nine SPIFI domains or the sum of all 82 indicators used to score each domain. Average scores for each domain are also shown, with all domain scores ranging from 1 to 7. In addition, Table 1 shows *p*-values from *t* tests comparing Waves 1 and 2 of data collection. Across all programs and waves, the sum of domain scores averages 15.8 of 63 (i.e., seven possible points across nine domains) and the sum of indicators averages 19.9 of 82. Average SPIFI scores increase between Waves 1 and 2 of data collection as programs received training and coaching in implementing the Pyramid Model. The differences between Waves 1 and 2 are statistically significant for the overall SPIFI scores and for individual domain scores except Domain 5 (behavior support plan) and Domain 8 (family engagement in individual children). These results, along with measures that assessed other aspects of Pyramid Model implementation, suggest that the SPIFI detects increasing implementation fidelity over time.

### Scale and Interrater Reliability

Table 2 displays measures of internal consistency (Cronbach's $\alpha$) and IRR. We used Cronbach's alpha to assess internal consistency, or scale reliability, among the nine domains, 82 indicators, and the subsets of indicators within each domain. The values of alpha ranged between .80 and .98, suggesting acceptable to high levels of internal consistency.

To assess IRR, a randomly selected subset of programs were each observed and scored by two evaluators as described above. To test IRR, we computed intraclass correlations (ICCs; Gamer et al., 2010; Hallgren, 2012) from multilevel models where evaluator observations are nested within programs. These models generally provide two variance estimates: the variance of errors associated with differences between evaluators within programs and the variation across programs. The ICC is the proportion of observed variance that occurs between programs; therefore, the ICC is high when variation between programs is large relative to variation among evaluators observing the same program.

As only a subset of programs were rated by multiple evaluators and we wanted to generalize the reliability of

**Table 2.** Reliability and Interrater Reliability ICC for Overall SPIFI and for Each Indicator Across Domains.

| Measure & domain | Cronbach's α | Interrater reliability (ICC, one-way model) |
|---|---|---|
| Overall SPIFI scores | | |
| Sum of all domains | .84 | .96 |
| Sum of all indicators | .98 | .98 |
| Item scores per domain | | |
| Domain 1: Leadership team | .89 | .86 |
| Domain 2: Leadership team activities | .95 | .89 |
| Domain 3: Staff buy-in | .91 | .83 |
| Domain 4: Program-wide expectations | .92 | .97 |
| Domain 5: Behavior support plan | .87 | .51 |
| Domain 6: Staff support plan | .88 | .96 |
| Domain 7: Family engagement | .85 | .98 |
| Domain 8: Family engagement individual children | .86 | .77 |
| Domain 9: Data-based decision-making | .80 | .85 |

*Note.* Single-measure ICCs are estimated from a one-way model. High rater reliability is characterized by absolute rater agreement, but ICCs are similar when consistency is examined instead of absolute agreement. SPIFI = Supporting Program-wide Implementation Fidelity Instrument; ICCs = intraclass correlations.

their ratings to programs rated by only one rater, we used a single-measure ICC. We also compared whether our ICCs differed when IRR was characterized by absolute agreement compared with consistency (i.e., evaluators providing scores with similar rank order). We found that the two estimates were nearly identical and report ICCs where IRR was based on absolute agreement between evaluators. Table 2 shows the ICC estimates for SPIFI scores calculated from the sum of all domains, the sum of all indicators, and for each domain score. Guidelines for interpreting these ICC estimates are provided by Landis and Koch (1977): poor < .00; slight: .00 to .20; fair: .21 to .40; moderate: .41 to .60; substantial: .61 to .80; almost perfect rater reliability: .81 to 1.00. Except for one indicator, all ICC estimates exhibit almost perfect or substantial reliability. Only Domain 5, behavior support plan, exhibited moderate IRR (ICC = .51).

## *Exploratory Factor Analyses*

To examine the latent structure of the SPIFI, we used EFA models, estimated using the principal factor method with an oblique rotation. We chose an oblique rotation under the theory that latent factors relevant to fidelity of implementation are likely correlated. Factor models from our EFA are estimated from the nine domain scores across all programs. To maximize sample size, we pooled all SPIFI scores from both waves of data collection. The scree plot suggests a possible one- or two-factor solution; therefore, we implemented parallel analysis to select the number of factors (Hayton et al., 2004; Horn, 1965). Parallel analysis allowed us to compare the scree plot from our observed data with data randomly generated to have the same number of indicators and programs. The intersection between the observed and generated plots indicates the optimal number of factors. Our plots cross between two and three factors, providing evidence that a two-factor model better represents the latent factor structure than a unidimensional, one-factor model.

We then examined measures of model fit to compare the one-factor and two-factor solutions. The two-factor solution achieves moderately good fit (standardized root mean square residual [SRMR] = 0.066, root mean square error of approximation [RMSEA] = .114 with 90% confidence interval [CI] = [0.070, 0.159], $\chi^2(19) = 43.681$, $p = .001$, comparative fit index [CFI] = .934, Tucker–Lewis index [TLI] = .874) that is better than the one-factor model (SRMR = 0.179, RMSEA = .188 with 90% CI = [0.156, 0.222], $\chi^2(28) = 127.209$, $p = .000$, CFI = .734, TLI = .658). These analyses provided suggestive evidence supporting a two-factor structure and we urge future research to continue exploring the latent factor structure of the SPIFI with larger sample sizes.

Factor loadings also provided evidence that indicators did not all load well onto one factor. Table 3 shows the factor loadings and uniqueness values from the two-factor model. The uniqueness values show the percentage of variance in the indicator score that is not explained by common factors. The first factor had larger loadings on Domain 1 (leadership team), Domain 2 (leadership team activities), Domain 3 (staff buy-in), and Domain 6 (staff support plan). The second factor had larger loadings on Domain 4 (program-wide expectations), Domain 5 (behavioral support plan), Domain 7 (family engagement), Domain 8 (family engagement for individual children), and Domain 9 (data-based decision-making). The first factor is more related to how program leaders encourage and sustain Pyramid Model implementation, which we labeled *Leadership*

**Table 3.** Factor Loadings and Uniqueness From Exploratory Factor Analysis.

| Indicators across domains | Factor 1 | Factor 2 | Uniqueness |
|---|---|---|---|
| Domain 1: Leadership team | 0.63 | 0.22 | 0.46 |
| Domain 2: Leadership team activities | 0.82 | −0.09 | 0.37 |
| Domain 3: Staff buy-in | 0.97 | −0.09 | 0.12 |
| Domain 4: Program-wide expectations | 0.21 | 0.79 | 0.22 |
| Domain 5: Behavior support plan | 0.02 | 0.69 | 0.52 |
| Domain 6: Staff support plan | 0.80 | 0.10 | 0.30 |
| Domain 7: Family engagement | −0.12 | 0.72 | 0.53 |
| Domain 8: Family engagement individual children | 0.04 | 0.59 | 0.64 |
| Domain 9: Data-based decision-making | −0.02 | 0.57 | 0.68 |

*Note.* Factors are based off oblique (oblimin) rotation. Factor 1 = Leadership Implementation Support; Factor 2 = Programmatic Implementation Support.

*Implementation Support*. Loadings for the second factor are more relevant to how the program operated as a cohesive unit, which we labeled *Programmatic Implementation Support*. Using results from the EFA, we predicted factor scores for both of these latent variables and standardized them to have a mean of 0 and standard deviation of 1.

For the remainder of this analysis, we report psychometric results on the sum of all domains, the sum of all indicators, the two predicted factor scores from the EFA, and each of the nine domain scores for a total 13 types of SPIFI scores. Summing the domains and the indicators is a scoring method that implicitly assumes unidimensionality, but we continued investigating these measures despite the weaker model fit of a one-factor model because the reliability of the overall measures and the ability to detect change over time along with the simplicity of these methods of aggregating the SPIFI may make the overall scores preferable to program personnel if the convergent validity is high. Therefore, it is important to understand the association between these measures and other measures of implementation fidelity. Finally, both overall scoring methods are highly correlated with each other and with both of the predicted latent factors. For example, the correlation between the sums of all SPIFI domains and all SPIFI indicators and both the administrative implementation support and organizational implementation support factors are between .77 and .96. These high correlations suggest that substantive conclusions are unlikely to be affected if researchers use one of the two unidimensional scoring methods.

## Convergent Validity

To assess convergent validity, we examined the relation between SPIFI scores and an alternative program-level implementation measure (PreSET; Steed et al., 2012), a classroom-level measure of Pyramid Model implementation (TPOT; Fox et al., 2014), and a measure of high quality classroom interactional practices (CLASS; Pianta et al., 2008). In selecting these measures, we expected stronger correlations of SPIFI with PreSET as they were designed to measure features of program implementation of similar approaches and with TPOT as a measure of the classroom practices that were the focus of program-wide implementation. We expected to see a relation between the SPIFI and CLASS as the CLASS measures the interactional and instructional quality of a classroom and includes dimensions measuring emotional climate, classroom organization, and instructional supports. Table 4 shows how SPIFI scores are correlated with the PreSET score, the TPOT score, and the three domains of the CLASS. In these analyses, all measures have been standardized to have a mean of 0 and a standard deviation of 1. The correlations show that SPIFI scores are most strongly correlated with the PreSET score (.67–.85). These higher correlations reflect a moderate to high degree of consistency between SPIFI and PreSET as program-level measures of implementation. The Programmatic Implementation Support factor correlated more highly with the PreSET than the Leadership Implementation Support but both are strongly correlated. Compared with the PreSET, correlations with the TPOT and each of the CLASS domains are smaller in magnitude but all positive and statistically significant, ranging from .23 to .47.

Then, we formally tested the association between SPIFI scores and measures of implementation or classroom practices (i.e., PreSET, TPOT, and CLASS) using two-level hierarchical linear models (HLMs) with classrooms (Level 1) nested within programs (Level 2). The HLMs estimate random intercepts for each program resulting in separate variance terms at the classroom and program level to account for the nesting of classrooms within programs (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). The HLM models are especially important in this context because SPIFI measures are collected at the program level. Each HLM model separately includes one of the 13 types of SPIFI scores as the focal predictor of interest. Repeating this process results in 65 separate mixed effect regression

**Table 4.** Correlation Between SPIFI Scores and External Measures.

| SPIFI Measurement and Domains | PreSET score | TPOT score | CLASS | | |
|---|---|---|---|---|---|
| | | | Emotional support | Classroom organization | Instructional support |
| Sum of SPIFI domains | .80* | .42* | .32* | .34* | .45* |
| Sum of SPIFI indicators | .83* | .44* | .32* | .34* | .42* |
| Leadership implementation support | .67* | .34* | .23* | .25* | .34* |
| Programmatic implementation support | .85* | .47* | .40* | .40* | .51* |

*Note.* All measures have been standardized to have a mean of 0 and a standard deviation of 1. SPIFI = Supporting Program-wide Implementation Fidelity Instrument; PreSET = Preschool-wide Evaluation Tool; TPOT = Teaching Pyramid Observation Tool; CLASS = Classroom Assessment Scoring System.
*$p < .05$ using Bonferroni-adjusted significance levels.

models (13 SPIFI measures × five measures of implementation or classroom practices). Without covariates, the HLM model for one outcome measure $y$ (e.g., PreSET, TPOT, or CLASS domain) for classroom $i$ in program $j$ is of the form:

$$\begin{aligned} &\text{Level 1 model:} \quad y_{ij} = \beta_{0j} + e_{ij}, \\ &\text{Level 2 model:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}\text{SPIFI}_j + u_{oj}, \\ &\text{Reduced form:} \quad y_{ij} = \gamma_{00} + \gamma_{01}\text{SPIFI}_j + u_{0j} + e_{ij}. \end{aligned}$$

The model includes a constant $\gamma_{00}$ representing the average value of $y$ for programs receiving a standardized SPIFI score of zero, a random intercept for program, $u_{0j}$, and a classroom level error term, $e_{ij}$. The coefficient of interest is $\gamma_{01}$. As all measures have been standardized, $\gamma_{01}$ is interpreted as a change in $y$ (in standard deviations) associated with a one standard deviation unit increase on the relevant SPIFI score.

To help rule out alternative explanations for the relations estimated by the HLM models and improve precision of the estimates, covariates at both the classroom and program level were also added to the model shown above. Classroom covariates include the teacher's teaching experience (in months), whether the teacher has a relevant degree in early childhood education, and teacher's race. Program covariates include the number of children served by the program, the number of administrators, whether the program is a Head Start program, if the program currently implements a social emotional learning curriculum, the proportion of female children, proportions by child race (with White as the reference category), the proportion of children with an IEP or IFSP, and the proportion of children who are dual language learners. To maximize our sample, we pooled together data from the first and second waves of data collection and include a wave indicator to control for differences between the two waves. We also tested these HLM models for each wave separately and reached substantively similar conclusions.

Table 5 shows the results from our two-level HLM models. Note that each reported coefficient comes from a separate regression of each outcome measure (labeled in the columns) on a separate SPIFI score. The first column shows the relation between each SPIFI score and the PreSET score. For example, a one standard deviation increase in the sum of the SPIFI domain is associated with a 0.753 standard deviation increase in the PreSET score, holding constant all classroom- and program-level covariates. The first column provides evidence of convergent validity because the PreSET is a validated measure of implementation fidelity for positive behavior supports (Steed & Webb, 2013) and the association between SPIFI and PreSET is positive and statistically significant for all of the various SPIFI scoring methods and each of the nine SPIFI items. Column 2 shows that the SPIFI is also correlated with the TPOT, a measure of Pyramid Model implementation at the classroom level. The relations with SPIFI items are statistically significant except for Item 5 (behavior support plan) and Item 8 (family engagement for individual children). The strength of the relationships between SPIFI measures and PreSET is also stronger than the association between SPIFI and TPOT. All four aggregated measures of the SPIFI as well as three individual items (leadership team [Item 1], staff buy-in [Item 3], and family engagement for individual children [Item 8]) are positive and significantly correlated with the Emotional Support dimension of the CLASS. The Classroom Organization dimension of the CLASS did not have a statistically significant association with overall SPIFI scores but did have small and statistically significant correlations to staff buy-in (Item 3), program-wide expectations (Item 4), and family engagement (Item 7). The instructional support dimension of CLASS only has a small and statistically significant correlation with program-wide expectations (Item 4).

## Validity Check

A second sample of SPIFI data was collected by 27 external coaches across the United States and Canada who were engaged in supporting Pyramid Model implementation in their state or province as a means of providing a robustness

**Table 5.** Standardized Coefficient Estimates From Two-Level Mixed Models Using SPIFI Scores as Predictors.

| | | | CLASS | | |
|---|---|---|---|---|---|
| SPIFI Measurement and Domains | PreSET score | TPOT score | Emotional support | Classroom organization | Instructional support |
| Sum of SPIFI domains | 0.753*** | 0.387*** | 0.272* | 0.177 | 0.152 |
| | (0.038) | (0.095) | (0.11) | (0.10) | (0.095) |
| Sum of SPIFI indicators | 0.730*** | 0.424*** | 0.235* | 0.203 | 0.165 |
| | (0.032) | (0.098) | (0.12) | (0.11) | (0.099) |
| Leadership implementation support | 0.739*** | 0.470*** | 0.260* | 0.211 | 0.175 |
| | (0.048) | (0.099) | (0.12) | (0.11) | (0.10) |
| Programmatic implementation support | 0.826*** | 0.330** | 0.256* | 0.169 | 0.141 |
| | (0.019) | (0.10) | (0.12) | (0.11) | (0.10) |
| Domain 1: Leadership team | 0.330*** | 0.362*** | 0.338*** | 0.121 | 0.0917 |
| | (0.073) | (0.088) | (0.099) | (0.096) | (0.089) |
| Domain 2: Leadership team activities | 0.595*** | 0.334** | 0.179 | 0.0984 | 0.137 |
| | (0.073) | (0.12) | (0.13) | (0.12) | (0.11) |
| Domain 3: Staff buy-in | 0.574*** | 0.478*** | 0.348** | 0.211* | 0.179 |
| | (0.063) | (0.096) | (0.11) | (0.11) | (0.099) |
| Domain 4: Program-wide expectations | 0.864*** | 0.505*** | 0.234 | 0.293* | 0.225* |
| | (0.048) | (0.11) | (0.13) | (0.12) | (0.11) |
| Domain 5: Behavior support plan | 0.961*** | 0.166 | 0.0309 | −0.0483 | 0.0211 |
| | (0.044) | (0.12) | (0.13) | (0.12) | (0.11) |
| Domain 6: Staff support plan | 0.753*** | 0.385*** | 0.0169 | 0.154 | 0.133 |
| | (0.031) | (0.094) | (0.12) | (0.10) | (0.094) |
| Domain 7: Family engagement | 0.649*** | 0.303** | 0.0601 | 0.209* | 0.149 |
| | (0.051) | (0.093) | (0.11) | (0.096) | (0.090) |
| Domain 8: Family engagement in individual children | 0.282** | 0.0311 | 0.308* | 0.125 | 0.101 |
| | (0.099) | (0.12) | (0.12) | (0.12) | (0.11) |
| Domain 9: Data-based decision-making | 0.517*** | 0.315*** | 0.188 | 0.111 | 0.0600 |
| | (0.045) | (0.093) | (0.11) | (0.098) | (0.091) |
| Covariates | Yes | Yes | Yes | Yes | Yes |

*Note.* Each cell shows the standardized coefficient for an individual multilevel model. Standard errors in parentheses. All measures are standardized to have a mean of 0 and a standard deviation of 1. SPIFI = Supporting Program-wide Implementation Fidelity Instrument; PreSET = Preschool-wide Evaluation Tool; TPOT = Teaching Pyramid Observation Tool; CLASS = Classroom Assessment Scoring System.
*$p < .05$. **$p < .01$. ***$p < .001$.

check for psychometric properties of the SPIFI using a second study sample and different procedures for training evaluators. We examined the robustness of the tool when used in the field by external coaches who worked a variety of implementation coaches. The notion of robustness refers to the ability of the tool to provide a meaningful measure within an application where some user errors might occur. The 17 programs in the second sample were located in California, Florida, Iowa, Minnesota, Pennsylvania, and Wisconsin in the United States and Nova Scotia in Canada. Programs ranged in size from two to 16 classrooms and 23 to 435 children. Ten programs provided tuition or fee-based child care, two provided Head Start or Early Head Start services, and nine provided public preschool services.

The external coaches for programs in the second sample collected SPIFI data for this robustness check. These external coaches were providing coaching to a local early childhood program that had a leadership team and had been engaging in Pyramid Model implementation for at least 1 year. Whenever possible, we asked two participating external coaches to conduct the SPIFI with one program together—one designated as a primary and one as the reliability evaluator. These coaches participated in a 2-hr synchronous online training on procedures for data collection and scoring with the tool and scoring guide. Data collectors in both the national sample and pilot study followed the same protocol when using the SPIFI. The data were collected concurrently with the data collection in the pilot sites.

The overall SPIFI scores and all of the domain scores are lower among the second wave of the pilot study programs than the national sample programs at statistically significantly levels with $p$ values lower than .01 (see the online Supplemental Tables 1, 2, and 3), likely due to a longer period of implementation within the national sample. The psychometric properties of the SPIFI data from the pilot and

robustness samples are largely consistent. The reliabilities (Cronbach's α) are high but somewhat lower in magnitude in the robustness sample. The interrater reliabilities are lower for the robustness sample but high to moderate but for three of the domain scores. The pattern of higher and lower loadings from each of the domains on the two factors is similar across the two samples, except for the relatively low loading of the Leadership Team domain on Factor 1 for the robustness sample.

## Discussion

We sought to validate a measure that would provide an objective assessment of the Pyramid Model implementation in early childhood programs. Our goal was to have a reliable and sensitive tool that could measure the level of implementation of critical features and use these data to provide technical assistance and stronger implementation fidelity. We used a multiphase approach to develop the SPIFI that incorporated feedback from experts and professionals in the field. We used data from a pilot study of PWS-PMI to examine the psychometric properties of the instrument. The data were evaluated by our research team to examine the distributions, reliability, and IRR. We then used EFA to determine the structure of the domains and HLM to examine the relationship between the overall SPIFI measures and each of the nine domains and other validated instruments (e.g., TPOT, PreSET) that measure features relevant to Pyramid Model implementation.

Our findings from this psychometric integrity study of the SPIFI indicate that it validly and reliably measures the program-level fidelity of implementation of the Pyramid Model. The EFA provided evidence that the overall SPIFI has two latent constructs: (a) leadership implementation support and (b) programmatic implementation support. The instrument also detected differences in implementation levels between Wave 1 and Wave 2 in pilot study programs, which indicates that SPIFI is sufficiently sensitive to programmatic changes to detect change over time in the same programs. In addition, implementation levels in Wave 1 of the pilot sample were significantly lower than more experienced programs in the robustness sample.

The data also revealed a correlation between the SPIFI and PreSET, both of which measure programmatic supports. The SPIFI's emphasis, however, on specific systemic factors that facilitate Pyramid Model and less so on classroom functioning, is reflected in the sources of data and the resulting factor loading. The stronger association between the SPIFI and PreSET versus the SPIFI and TPOT suggests that the classroom-level implementation fidelity measure is not a perfect indicator of program-level fidelity and the two should be measured separately. The correlation with TPOT scores points to a relation between program-level implementation and classroom practices. The consistency of the results of the factor analysis and relation between other measures and the SPIFI measures, both aggregate and by domain, suggests that the SPIFI tools are robust to differences in samples, training of evaluators, and experience with implementing the Pyramid Model. Overall, these results support the validity of the SPIFI as a program-wide instrument for measuring the fidelity of Pyramid Model implementation.

### *Limitations and Future Research*

There are several limitations with the study. For example, our sample size was relatively small. A larger and more diverse sample should be used in future replications to evaluate SPIFI and program-wide implementation of the Pyramid Model. With data on program and child performance, more nuanced analyses can examine the sensitivity of the instrument in differentiating programs in exploratory phases of implementation and those moving to full implementation. A larger sample also would allow examinations of teacher and staff backgrounds (e.g., certification) as they relate to systemic implementation. Given four domains that loaded similarly onto the two latent factors, future research including more programs would allow researchers to better examine whether these four domains could be modified to better discriminate between the two factors or whether the data would better fit a different latent structure. Using a larger sample with raters trained in the same way would ensure that any observed differences in the SPIFI scores are not driven by differences in raters, as may be the case in this study. We noted that IRR was lower in the data-based decision-making domain. This could be due to the lower overall implementation in this area after 1 year and, therefore, less access to data for scoring purposes. Evaluation of programs that are further in the implementation process might clarify this issue. Future research might examine the extent to which familiarity with a program affects the evaluation process and outcomes. Our study demonstrates the use of the SPIFI by typical evaluators and provides initial validation of the SPIFI as a measure of implementation fidelity that can be used in evaluative research and technical assistance.

### Acknowledgments

The authors wish to acknowledge the programs, teachers, and children who participated in the study, the administrators who supported the study, and our staff and students who supported this work.

### Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: M.L.H. and L.F. are authors of the Teaching Pyramid Observation Tool and receive a portion of royalties.

## Funding

This work was supported by a grant from the National Center for Education Research, Institute of Education Sciences, U.S. Department of Education to Vanderbilt University (R305A150141). The opinions expressed are those of the authors, not the funding agency.

## Supplemental Material

Supplementary material for this article is available on the *Journal of Positive Behavior Interventions* website with the online version of this article.

## References

Blair, K.-S. C., Fox, L., & Lentini, R. (2010). Use of positive behavior support to address the challenging behavior of young children within a community early childhood program. *Topics in Early Childhood Special Education*, *30*(2), 68–79. https://doi.org/10.1177/0271121410372676

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, *25*(2), 166–176. https://doi.org/10.1016/j.ecresq.2009.10.004

Domitrovich, C. E., Moore, J. E., & Greenberg, M. T. (2012). Maximizing the effectiveness of social-emotional interventions for young children through high-quality implementation of evidence-based interventions. In B. Kelly & D. F. Perkins (Eds.), *Handbook of implementation science for psychology in education* (pp. 207–229). Cambridge University Press. https://doi.org/10.1017/CBO9781139013949.017

Durlak, J. A. (2016). Programme implementation in social and emotional learning: Basic issues and research findings. *Cambridge Journal of Education*, *46*, 333–345. https://doi.org/10.1080/0305764X.2016.1142504

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531–540. https://doi.org/10.1177/1049731509335549

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication #231). The National Implementation Research Network, University of South Florida.

Fox, L., Dunlap, G., Hemmeter, M. L., Joseph, G. E., & Strain, P. S. (2003). The teaching pyramid: A model for supporting social competence and preventing challenging behavior in young children. *Young Children*, *58*(4), 48–52. https://doi.org/10.1002/cbl.20134

Fox, L., & Hemmeter, M. L. (2009). A programwide model for supporting social emotional development and addressing challenging behavior in early childhood settings. In *Handbook of positive behavior support* (pp. 177–202). Springer.

Fox, L., Hemmeter, M. L., Jack, S., & Binder, D. (2017). *Early childhood program-wide PBS benchmarks of quality*. National Center for Pyramid Model Innovations.

Fox, L., Hemmeter, M. L., & Snyder, P. S. (2014). *Teaching Pyramid Observation Tool for Preschool Classrooms (TPOT™), research edition*. Paul H. Brookes.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2010). *Various coefficients of interrater reliability and agreement* (R package version 0.83). https://www.researchgate.net/publication/260283026_irr_Various_Coefficients_of_Interrater_Reliability_and_Agreement

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Hemmeter, M. L., Barton, E., Fox, L., Vatland, C., Henry, G., Pham, L., Horth, K., Taylor, A., Binder, D. P., von der Embse, M., & Veguilla, M. (2022). Program-wide implementation of the Pyramid Model: Supporting fidelity at the program and classroom levels. *Early Childhood Research Quarterly*, *59*, 56–73. https://doi.org/10.1016/j.ecresq.2021.10.003

Hemmeter, M. L., Fox, L., & Snyder, P. (2013). A tiered model for promoting social-emotional competence and addressing challenging behavior. In V. Buysse & E. Peisner-Feinberg (Eds.), *Handbook of response to intervention in early childhood* (pp. 85–101). Brookes.

Hemmeter, M. L., Fox, L., Snyder, P., Algina, J., Hardy, J. K., Bishop, C., & Veguilla, M. (2021). Corollary child outcomes from the Pyramid Model professional development intervention efficacy trial. *Early Childhood Research Quarterly*, *54*, 204–218.

Hemmeter, M. L., Snyder, P. A., Fox, L., & Algina, J. (2016). Evaluating the implementation of the Pyramid Model for Promoting Social-Emotional Competence in early childhood classrooms. *Topics in Early Childhood Special Education*, *36*(3), 133–146. https://doi.org/10.1177/0271121416653386

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Meek, S. E., & Gilliam, W. S. (2016). *Expulsion and suspension as matters of social justice and health equity* [Discussion Paper]. National Academy of Medicine.

Metz, A., Bartley, L., Ball, H., Wilson, D., Naoom, S., & Redmond, P. (2015). Active implementation frameworks for successful service delivery: Catawba county child wellbeing project. *Research on Social Work Practice*, *25*(4), 415–422. https://doi.org/10.1177/1049731514543667

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™ (CLASS™) Manual: Pre-K*. Paul H. Brookes Publishing.

Quesenberry, A. C., Hemmeter, M. L., & Ostrosky, M. M. (2011). Addressing challenging behaviors in Head Start: A closer look at program policies and procedures. *Topics in Early Childhood Special Education*, *30*(4), 209–220. https://doi.org/10.1177/0271121410371985

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). SAGE.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. SAGE.

Snijders, T., & Bosker, R. (2012). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

Snyder, P. A., Hemmeter, M. L., & Fox, L. (2022). *Essentials of practice-based coaching. Supporting effective practices in early childhood*. Paul H. Brookes.

Snyder, P. A., Hemmeter, M. L., Fox, L., Bishop, C. C., & Miller, M. D. (2013). Developing and gathering psychometric evidence for a fidelity instrument: The Teaching Pyramid Observation Tool–Pilot version. *Journal of Early Intervention*, *35*(2), 150–172. https://doi.org/10.1177/1053815113516794

StataCorp. (2017). *Stata Statistical Software: Release 15*.

Steed, E. A., Pomerleau, T., & Horner, R. H. (2012). *Preschool-wide evaluation tool (PreSET)*. Paul H. Brookes.

Steed, E. A., & Webb, M. Y. L. (2013). The Psychometric Properties of the Preschool-Wide Evaluation Tool (PreSET). *Journal of Positive Behavior Interventions*, *15*(4), 231–241. https://doi.org/10.1177/1098300712459357

Sugai, G., & Horner, R. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child & Family Behavior Therapy*, *24*(1–2), 23–50. https://doi.org/10.1300/J019v24n01_03

U.S. Department of Health and Human Services and U.S. Department of Education. (2014). *Policy statement on expulsion and suspension policies in early childhood settings*. https://www.acf.hhs.gov/sites/default/files/ecd/expulsion_suspension_final.pdf

Wasik, B. A., Mattera, S. K., Lloyd, C. M., & Boller, K. (2013). *Intervention dosage in early childhood care and education: It's complicated* (OPRE Research Brief OPRE 2013-15). Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE.