Journal of Early Intervention
Volume XX Number X
Month XXXX xx-xx

10.1177/1053815113516794
http://jei.sagepub.com
hosted at
http://online.sagepub.com

Developing and Gathering Psychometric Evidence for a Fidelity Instrument

The Teaching Pyramid Observation Tool–Pilot Version

Patricia A. Snyder
University of Florida, Gainesville, USA
Mary Louise Hemmeter
Vanderbilt University, Nashville, USA
Lise Fox
University of South Florida, Tampa, USA
Crystal Crowe Bishop
M. David Miller
University of Florida, Gainesville, USA

Fidelity assessment has received renewed attention in recent years, particularly as distinctions have been made in implementation science between intervention fidelity and implementation fidelity. Considering both types of fidelity has been recommended when developing fidelity instruments. In the present article, we describe development of the pilot version of the Teaching Pyramid Observation Tool (TPOT-P) as a case example of designing a fidelity instrument for use in research and practice. The TPOT is a multimethod judgment-based rating scale designed to measure practitioners' fidelity of implementation of practices associated with the *Pyramid Model*. We describe the structure of the TPOT-P in relation to *Pyramid Model* components and fidelity indicators. We summarize the measurement approaches grounded in generalizability theory and classical test theory that were used to investigate the psychometric properties of TPOT-P scores based on data collected by trained raters on three occasions in 50 preschool classrooms. Findings suggest the TPOT-P shows promise for dependably measuring teachers' implementation of *Pyramid Model* practices.

Keywords: *implementation fidelity, assessment, social-emotional teaching practices, generalizability theory, preschool*

Authors' Note: Patricia A. Snyder, School of Special Education, School Psychology, and Early Childhood Studies, University of Florida; Mary Louise Hemmeter, Department of Special Education, Peabody College of Education and Human Development, Vanderbilt University; Lise Fox, College of Behavioral and Community Sciences, University of South Florida; Crystal Crowe Bishop, School of Special Education, School Psychology, and Early Childhood Studies, University of Florida; M. David Miller, School of Human Development and Organizational Studies in Education, University of Florida. Work reported in this article was supported, in part, by a grant from the National Center for Special Education Research in the Institute of Education Sciences to Vanderbilt University (R324A07212). The opinions expressed are those of the authors, not the funding agency. Correspondence regarding this article should be directed to Patricia Snyder, School of Special Education, School Psychology, and Early Childhood Studies, University of Florida, Gainesville, FL 32611; email: patriciasnyder@coe.ufl.edu

Fidelity assessment has traditionally been described as processes or procedures used to gather information about the extent to which components of a curriculum, intervention, or practice are implemented as intended. Assessment of fidelity has been identified as integral for linking teaching practices to observed child outcomes and for establishing the efficacy of interventions used in educational or early learning settings (Hagermoser Sanetti, Dobey, & Gritter, 2012; Horner et al., 2005; Lloyd, Supplee, & Mattera, 2013; O'Donnell, 2008).

The importance of fidelity assessment in research, which includes evaluating implementation of the independent variable and examining relationships to outcomes, has been recognized for many years in early intervention and early childhood special education (LeLaurin & Wolery, 1992, Wolery, 2011). In recent years, fidelity assessment has received renewed attention given growing interest in the science of implementation. Implementation science examines the supports and processes necessary for programs, practices, or interventions with promising research evidence to be implemented on a large scale and in authentic settings (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Halle, Metz, & Martinez-Beck, 2013).

As implementation science in early childhood has evolved, distinctions have been made between the assessment of *intervention fidelity* and the assessment of *implementation fidelity* and the importance of considering both when designing fidelity assessments (Downer, 2013; Hulleman, Rimm-Kaufman, & Abry, 2013; Lloyd et al., 2013). Intervention fidelity has been defined as the extent to which a program, curriculum, intervention, or practice is implemented as intended. Critical to the assessment of intervention fidelity is identifying the core components or active ingredients of a program, curriculum, intervention, or practice (Hulleman et al., 2013). O'Donnell (2008) described five aspects of fidelity that might be assessed: (a) adherence, the extent to which the intervention is delivered as intended; (b) duration, the number, length, and frequency of intervention sessions completed; (c) quality of delivery, the manner in which the components of the intervention are implemented; (d) participant responsiveness, the extent to which participants are involved in the intervention; and (e) program differentiation, whether components distinguishing one intervention from another are present in the implementation of the intervention. Intervention fidelity can be measured in a variety of ways, including direct observation, use of video or audio recordings, practitioner interviews, practitioner self-report surveys, and reviews of classroom documentation.

Implementation fidelity focuses on contextual factors or "drivers" that support implementation of the intervention and its core components (Downer, 2013). Metz and Bartley (2012) described three major classes of implementation fidelity drivers: (a) competency, (b) organization, and (c) leadership. Examples of competency drivers are staff selection, training, coaching, and performance assessment. Organization drivers are decision-support data systems, facilitative administration, and systems interventions. Leadership drivers include technical and adaptive leadership strategies (Metz & Bartley, 2012; Metz, Halle, Bartley, & Blasberg, 2013). Because the drivers support intervention implementation, assessment of the fidelity with which the drivers are implemented as intended is important. Nevertheless, implementation fidelity assessment is less commonly reported in the literature than intervention fidelity assessment (Downer & Yazejian, 2013). Implementation drivers usually are not identified as a key component of an intervention so intervention fidelity assessments typically do not include implementation components and assessment

of implementation. One exception would be a professional development intervention in which competency drivers such as training and coaching might be identified as key components and included as part of an assessment of intervention fidelity.

Recommendations have been made to better align the assessment of both types of fidelity early in the process of developing and evaluating interventions and to develop well-constructed, multimethod fidelity measures for which reliability and validity evidence is gathered and evaluated (Downer, 2013; Hulleman et al., 2013). As part of a development and innovation research project funded by the Institute of Education Sciences, a fidelity instrument was developed and preliminary psychometric integrity evidence for the instrument was gathered. The instrument would subsequently be used as a measure of intervention fidelity in a potential efficacy randomized controlled trial and might hold promise for assessment of implementation fidelity. The instrument was known as the *Teaching Pyramid Observation Tool–Pilot Version* (TPOT-P; Fox, Hemmeter, & Snyder, 2008).

The TPOT-P was developed and evaluated in tandem with the development and validation of the *Pyramid Model* intervention (Hemmeter, Fox, & Snyder, 2007), which is a professional development intervention designed to support teachers' implementation of practices associated with a tiered framework focused on social-emotional competence and behavior support. The components of the professional development intervention included the provision of intensive workshops (18 hr over 3 days); expert coaching following the workshops, provided in teachers' classrooms for approximately 13 weeks with each coaching session lasting approximately 90 min; and guides and materials to supplement the workshops and coaching and to further support teachers' implementation of *Pyramid Model* practices.

The TPOT-P was designed to support inferences about three aspects of intervention fidelity described by O'Donnell (2008): (a) adherence, (b) quality, and (c) differentiation. Its intended uses in our research project were to characterize intervention fidelity in baseline and treatment conditions as part of feasibility studies (Fox, Hemmeter, Snyder, Binder, & Clarke, 2011) and in intervention and counterfactual conditions in a potential efficacy trial as well to evaluate relationships between intervention fidelity and child outcomes (Hemmeter, Snyder, Fox, & Algina, 2011). In addition, the TPOT-P was designed to be used by trained coaches to provide support and feedback to intervention teachers involved in the feasibility study and potential efficacy trial about their implementation of *Pyramid Model* practices. With respect to assessment of implementation fidelity, the TPOT-P might be useful as a competency assessment instrument (e.g., for coaches to receive support and feedback about *Pyramid Model* practices on which they focus with teachers; for use in professional development and performance assessment systems).

We describe in this article the development of the TPOT-P as a case example of designing an instrument for use in assessing intervention and implementation fidelities. In addition, we present findings from a study conducted to gather preliminary psychometric integrity evidence for the instrument. To situate the development of the TPOT-P, we first describe the *Pyramid Model* and its associated components. We then describe the development of the TPOT-P and discuss how fidelity indicators were identified and operationalized to align with *Pyramid Model* components. Finally, we present findings from a study in which we used descriptive analyses and generalizability theory in conjunction with convergent score validity evidence to address three primary research questions:

Research Questions 1: What are the mean and range of TPOT-P scores in preschool classrooms that have not received an intervention focused on the *Pyramid Model* and associated practices?

Research Questions 2: How much variance in observed TPOT-P scores is associated with raters, TPOT-P indicators, and rating occasions?

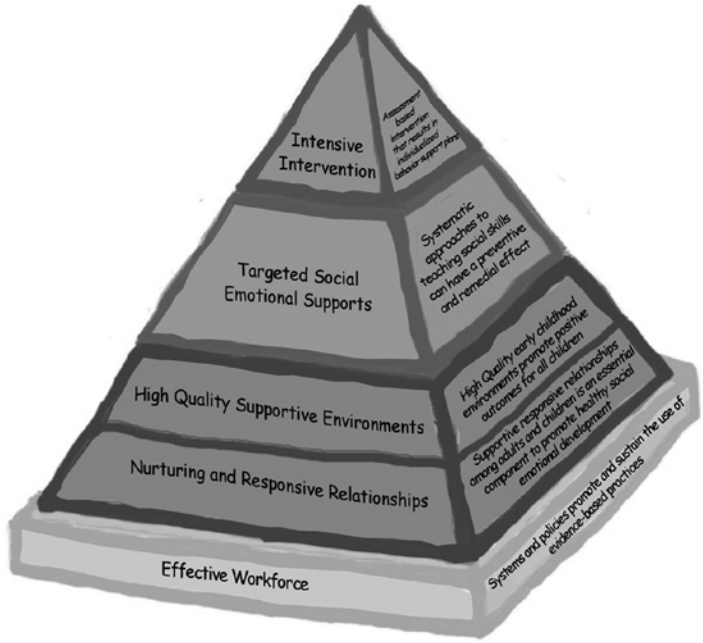
Research Questions 3: What is the convergent score validity evidence for the TPOT-P when the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre, 2008) is used to quantify observed interactions and use of curricular materials under three domains (emotional support, instructional support, and classroom organization)?

We explored the extent to which TPOT-P indicators were implemented by teachers who had not been systematically exposed to the *Pyramid Model* intervention to identify indicators that might differentiate implementation of *Pyramid Model* practices from implementation of other social-emotional practices in preparation for using the instrument in the feasibility study and potential efficacy trial (Bond, Evans, Salyers, Williams, & Kim, 2000; Mowbray, Holter, Teague, & Bybee, 2003). We also were interested in examining score dependability with raters, TPOT-P indicators, and rating occasions as facets of measurement prior to using the TPOT-P in a potential efficacy trial that would involve trained raters administering the instrument on multiple occasions in preschool classrooms. Finally, we explored relationships between TPOT-P scores and scores on the CLASS (Pianta et al., 2008). The CLASS was selected as an instrument for exploring convergent validity because its theoretical and empirical foundations emphasize key dimensions of classroom interactional processes, such as emotional and instructional support practices that contribute to classroom quality (LaParo, Pianta, & Stuhlman, 2004; Pianta et al., 2008). We hypothesized noteworthy associations between TPOT-P and CLASS scores, particularly between scores on the CLASS emotional and instructional support domains and key TPOT-P components.

Pyramid Model for Promoting Social-Emotional Competence in Infants and Young Children

The *Pyramid Model* is a tiered promotion, prevention, and intervention framework that organizes and guides the implementation of practices demonstrated to support children's acquisition and mastery of skills related to social-emotional competence as well as prevent or reduce challenging behavior (Fox, Carta, Strain, Dunlap, & Hemmeter, 2010; Fox, Dunlap, Hemmeter, Joseph, & Strain, 2003; Hemmeter, Fox, & Snyder, 2013; Hemmeter, Ostrosky, & Fox, 2006; Raver & Knitzer, 2002; Thompson & Goodman, 2009; Thompson & Raikes, 2007). As shown in Figure 1, the first tier of the *Pyramid Model* specifies two features of universal practices: (a) nurturing and responsive relationships and (b) high-quality supportive classroom environments. The second tier focuses on targeted practices related to explicit teaching of social and emotional skills, including skills to prevent or replace challenging behavior. The third tier focuses on practices related to individualizing social, emotional, and behavioral support interventions for children with significant social or emotional skill deficits and persistent challenging behavior.

Figure 1
Pyramid Model Framework



The practices associated with each tier of the *Pyramid Model* are based on research on effective instruction for young children (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; National Research Council, 2001), strategies to promote child engagement and appropriate behavior (Chien et al., 2010; Conroy, Brown, & Olive, 2008), the promotion of children's social skills (Brown, Odom, & McConnell, 2008; Vaughn et al., 2003), and the implementation of individualized assessment-based behavior support plans for children with the most severe behavior challenges (Blair, Fox, & Lentini, 2010; Conroy, Dunlap, Clarke, & Alter, 2005; Dunlap, Wilson, Strain, & Lee, 2013; McLaren & Nelson, 2009). Guidance for implementation and professional development resources related to the *Pyramid Model* and associated practices have been provided as part of two national centers, the Center on the Social and Emotional Foundations for Early Learning (CSEFEL) and the Technical Assistance Center on Social-Emotional Interventions (TACSEI).

Preschool Practices Associated With the *Pyramid Model*

The *Pyramid Model* includes practices appropriate for infant/toddler and preschool settings. Practice components of the *Pyramid Model* and associated fidelity indicators specified on the TPOT-P were those identified as appropriate for preschool classrooms. As

shown in Figure 1, the universal level of the *Pyramid Model* includes interactional and environmental practices identified as foundational for promoting skills related to social-emotional competence and positive behavior. Interactional practices focused on nurturing and responsive relationships emanate from a substantial body of empirical evidence that these relationships are pivotal to young children's development and learning (National Research Council, 2001; Shonkoff, 2010; Shonkoff & Phillips, 2000). Practices associated with relationships include supporting children's play and engagement, having supportive conversations with children, providing positive feedback and encouragement to children, and building positive relationships with children, families, and colleagues.

High-quality and supportive environments are considered a universal foundation for early development and learning (Mashburn & Pianta, 2010; Mashburn et al., 2008). For example, structural and temporal features of early childhood classrooms create the conditions in which teacher–child and child–child social interactions occur. This second set of universal practices associated with the *Pyramid Model* includes the provision of a predictable and supportive learning environment that maximizes the engagement of children within and across classroom activities and routines. Among the practices associated with high-quality and supportive environments are the provision of adequate materials, well-defined play and activity centers, balanced schedules, organized transitions, teaching about classroom routines, providing clear directions to children, offering positive and explicit guidance about rules and expectations, and designing activities that maximize child engagement and learning.

Secondary prevention practices focus on the provision of targeted social-emotional supports, including explicit instruction on social skills and emotional regulation (e.g., Denham et al., 2003; Domitrovich, Cortes, & Greenberg, 2007; Dunlap et al., 2003; McClelland, Morrison, & Holmes, 2000; Strain & Joseph, 2006; Webster-Stratton, 1999; Webster-Stratton & Reid, 2003). Targeted supports are offered to children who need additional guidance or intervention beyond that offered to all children. Social skills curricula might be implemented systematically with some children as part of secondary prevention in the *Pyramid Model* (Joseph & Strain, 2003). Practices associated with this level include teaching children how to identify and express emotions, problem-solve, and initiate and maintain interactions with adults and peers. Additional secondary prevention teaching practices are related to helping children learn how to handle disappointment and anger and develop friendships.

Tertiary practices in the *Pyramid Model* are individualized for children with persistent social-emotional competence skill deficits and challenging behavior. To inform decision making about which individualized practices to use with which children and under what circumstances, a team is convened. The team determines the nature and function of the social-emotional skill deficits or problem behavior and develops an individualized behavior support plan. The team implements the plan, conducts ongoing monitoring of child progress, and revises the plan, if needed (Fox & Hemmeter, 2009; Hemmeter et al., 2006). The individualized behavior support plan includes implementing prevention strategies to address "triggers" for challenging behavior, teaching replacement skills that are alternatives to the challenging behavior, and using strategies to reduce the occurrence of challenging behaviors (Dunlap et al., 2013; Fox et al., 2010). At this level, teaching and instructional practices are individualized and intensive for each child.

Development of the TPOT-P

The TPOT-P was developed for use in preschool classrooms. Fidelity indicators specified on the instrument were designed to assess a preschool teacher's classroom-wide implementation of universal and targeted teaching practices as well as a teacher's *capacity* to individualize teaching practices and implement individualized behavior support plans at the tertiary level. Due to the individualized nature of systematic instruction and individualized behavior support plans, the TPOT-P is not sufficient to use as a fidelity assessment of individualized instruction or implementation of individualized behavior support plans.

Development of an instrument that was a precursor to the TPOT-P began in 2005 with the goal of developing an efficient and practical tool that could be used in authentic preschool settings. As part of two federally funded projects (the Center for Evidence Based Practices: Young Children With Challenging Behavior funded by the Office of Special Education Programs and the CSEFEL funded by the Department of Health and Human Services), a comprehensive set of training materials was developed to introduce early childhood practitioners to the *Pyramid Model* and associated practices.

The CSEFEL training modules contained an *Inventory of Practices* that specified 130 practices associated with components of the *Pyramid Model*. This inventory was used as one source for identifying TPOT-P fidelity indicators. For organizing TPOT-P content in relation to the *Pyramid Model* and *Inventory of Practices*, three primary components were specified: (a) environmental arrangements, (b) key practices, and (c) red flags. The key practices component was divided into subcomponents (e.g., predictable schedules and routines, teaching children to express emotions). The following steps were used to develop TPOT-P content: (a) generated the list of components and subcomponents under which fidelity indicators could be organized based on a thorough review and synthesis of the literature (Dunlap et al., 2006; Joseph & Strain, 2003), (b) generated a definition for each component or subcomponent, (c) assigned practices from the *Inventory of Practices* to each component or subcomponent, (d) developed additional subcomponents when practices from the *Inventory of Practices* did not fit into one of the previously identified subcomponents, and (e) developed additional fidelity indicators not included on the *Inventory of Practices* but identified in the *Pyramid Model* as relevant for preschool classrooms.

Further development of the TPOT-P involved iterative processes of content validation by experts on the advisory boards associated with the two Centers and field-testing in authentic preschool settings as part of each Center's training and technical assistance activities. All preceding activities culminated in the pilot version of the TPOT, which was used in the present study.

Table 1 shows the three components included on the TPOT-P: (a) environmental arrangements, (b) key practices, and (c) red flags. In addition, 15 subcomponents associated with the key practices component are shown. Figure 2 shows an example of a key practices subcomponent (i.e., schedules and routines) and associated fidelity indicators. As illustrated in Table 1, the environmental arrangements component had 7 indicators. The 15 key practices subcomponents (e.g., schedules and routines, promoting children's engagement, teaching children to express emotions) each had between 4 and 10 fidelity indicators for a total of 118 (e.g., the subcomponent on schedules and routines had 9 indicators, the subcomponent on teaching children to express emotions had 8 indicators). The TPOT-P

Table 1
TPOT-P Components and Number of Fidelity Indicators Associated With Each Component

Component	Method[a]		No. of indicators
Environmental arrangements	O	Learning centers have clear boundaries (physical)	1
	O	The classroom is arranged such that all children in the classroom can move easily around the room	1
	O	The classroom is arranged such that there are no large, wide open spaces where children could run	1
	O	There is an adequate number and variety of centers of interest to children and to support the number of children (at least 4 centers; 1 center per every 4 children)	1
	O	Materials in all centers are adequate to support the number of children allowed to play	1
	O	Materials/centers are prepared before children arrive at the center or activity	1
	O	Classroom rules or program-wide expectations are posted, illustrated with a picture or photo of each rule or expectation, limited in number (3-5), and stated positively	1
		Total number of environmental arrangement indicators	7
Key practice subcomponents	O	Schedules and routines	9
	O	Transitions between activities	8
	O	Supportive conversations	10
	O	Promoting children's engagement	9
	O	Teaching children behavior expectations	7
	O	Providing directions	6
	O	Using effective strategies to respond to challenging behavior[b]	(10)
	O	Teaching social skills and emotional competencies	8
	O/I	Teaching children to express emotions	8
	O/I	Teaching problem solving	10
	O/I	Teaching friendship skills	9
	I	Supporting children with persistent problem behavior	4
	I	Communicating with families and promoting family involvement in the classroom	8
	I	Involving families in supporting their child's social-emotional development and addressing problem behavior	7
	I	Collaborative teaming relationships with other adults	5
		Total number of key practices indicators	108[c]
Red flags	O	The majority of the day is spent in teacher-directed activities	1
	O	Transitions are more often chaotic than not	1
	O	Teacher talk to children is primarily giving directions, telling children what to do, reprimanding children	1
	O	During group activities, many children are NOT engaged	1
	O	Teachers are not prepared for activities before the children arrive at the activity	1

(continued)

Table 1 (continued)

Component		Method[a]	No. of indicators
	O	Children are reprimanded for engaging in problem behavior (use of "no," "stop," "don't")	1
	O	Children are threatened with an impending negative consequence that will occur if problem behavior persists	1
	O	Teacher reprimands children for expressing their emotions	1
	O	Emotions are never discussed in the classroom	1
	O	Teacher rarely encourages interactions between children during play or activities	1
	O	Teacher gives directions to all children in the same way without giving additional help to children who need more support	1
	O	Teacher tells children mostly what not to do rather than what to do	1
	I	Teacher asks for the removal of children with persistent challenging behavior from the classroom or program	1
	I	Teacher comments about families are focused on the challenges presented by families and their lack of interest in being involved	1
	I	Teacher only communicates with families when children have challenging behavior	1
	I	Teacher complains about other team members and notes difficulty in their relationships	1
		Total number red flag indicators	16
		Total number of fidelity indicators on TPOT-P	131[d]

Note. *TPOT-P = Teaching Pyramid Observation Tool–Pilot.*
[a]Method used to inform scoring. O = observation, O/I = observation and interview. I = interview.
[b]Indicators associated with this component scored only if challenging behavior occurs during the TPOT-P observation. Due to the large amount of missing data associated with the scoring of this item, it was omitted from analyses in the present study.
[c]108 key practice component indicators scored on the TPOT-P if challenging behavior does not occur during the observation and 118 indicators scored if challenging behavior does occur.
[d]131 indicators scored on the TPOT-P if challenging behavior does not occur during the observation; 141 indicators scored if challenging behavior does occur.

also had a "red flag" component with 16 fidelity indicators. Red flags represented practices that were inconsistent or incompatible with implementation of *Pyramid Model* practices (e.g., teacher reprimands children for expressing emotions, transitions are more often chaotic than not). Across the environmental arrangements, key practices, and red flag components, there were 141 fidelity indicators on the TPOT-P.

Psychometric Integrity Study of the TPOT-P

Setting

Fifty preschool classrooms in middle Tennessee were part of the study. Thirty-seven (74%) were Head Start classrooms, 4 (8%) were inclusive early childhood special education

Figure 2
Example of key Practices Subcomponent: Schedules and Routines With Nine Fidelity Indicators

__Y __N 1. Teacher posts classroom schedule with visuals so that children are aware of the activity sequence of the day	__Y __N 4. Teacher reviews the schedule with children and refers to it throughout the day	__Y __N 8. Teacher only continues with a specific teacher-directed activity when the majority of children are actively engaged and interested
__Y __N 2. Teacher-directed activities are shorter than 20 minutes	__Y __N 5. Teacher structures routines so that there is a clear beginning, middle, and end	__Y __N 9. Individual children who need extra support are prepared for activities using an activity schedule or cues at the beginning of activities
__Y __N 3. There are both large- and small- group activities	__Y __N 6. There is a balance of child-directed and teacher-directed activities	
	__Y __N 7. If needed, teacher prepares children when changes are going to occur within the schedule (score N/O if no opportunity to observe)*	
	__N/O	

Note. Indicators are scored 0 (N or N/O, not present) or 1 (Y, present).

classrooms, and 9 (18%) were pre-K programs for at-risk children. The mean number of children enrolled in each classroom was 17.7 (range = 6-21).

Participants

Participants were 50 female preschool lead teachers in the 50 preschool classrooms. The mean age of the teachers was 37 years (*SD* = 11.5). Twenty-nine teachers identified their ethnicity as African American, 18 reported they were Caucasian, and 1 teacher reported she was Asian (2 participants did not provide information about ethnicity). All teachers were employed full-time. The majority of teachers (60%) reported having a bachelor's degree, 18% had a master's degree, 8% an associate's degree, 8% reported another type of degree (e.g., CDA credential, educational specialist), and 6% indicated having a high school degree.

The mean number of years of preschool teaching experience was 10.4 (*SD* = 8.3) and the mean number of years in their current classroom position was 6.7 (*SD* = 7.0). The mean number of children in each teacher's classroom with Individualized Educational Programs (IEPs) was 2 (*SD* = 2). Eighty-six percent of the teachers reported attending training in the past year on promoting social-emotional skills or addressing challenging behavior. Fifty percent of the teachers reported having children with persistent challenging behavior (defined as

challenging behaviors of significant intensity or duration that disrupt the child's or others' engagement and learning and persists over time). Of these teachers, the mean number of children in a classroom with persistent challenging behavior was reported as 3 (*SD* = 3).

Measures

TPOT-P. The content of the TPOT-P (Fox et al., 2008) was described previously. Administration of the TPOT-P included an observation and an interview. Observations were conducted for approximately 2 hr teacher-directed activities (e.g., large group circle, small group instruction), child-directed activities (e.g., center time, free play), and the transitions that occur between activities were observed. In addition, a 15- to 20-min structured interview with the teacher was conducted using the questions provided for certain key practice and red flag indicators. Table 1 shows the methods used to score indicators for each component or subcomponent.

When administering the TPOT-P, observers used a scoring form (Figure 2 is an excerpt from the scoring form). The form included instructions for completing the observation and interview, a place to note the start and ending time of the observation, a place to make notes about the children and adults present in the classroom during the observation, and a chart for recording the schedule of the classroom during the observation. Space for making notes during the observation and writing answers during the interview was also provided on the scoring form. Data collectors in the present study also had continuous access during their observations and interviews to a scoring manual with operational definitions and scoring criteria for each fidelity indicator (Fox et al., 2008).

With respect to scoring, TPOT-P fidelity indicators were scored as 1 (*present*) or 0 (*not present*) or *no opportunity* (the four indicators that could be scored *no opportunity* were scored as *not present* for data analyses). Red flag items were rated as 1 (*present*) or 0 (*not present*) and were reverse-scored for data analyses. Ten fidelity indicators for one of the key practices subcomponents (i.e., using effective strategies to respond to challenging behavior) were scored only if challenging behavior was seen during a TPOT-P observation. Due to missing data associated with this subcomponent and because challenging behavior occurred inconsistently across observations, we excluded it from the generalizability analyses reported in the present article. Across the remaining 14 key practices subcomponents, there were 108 fidelity indicators. A total TPOT-P score was calculated by summing the scores of the 7 environmental arrangement indicators, the indicators associated with the 14 TPOT-P key practices subcomponents, and the reverse-scored red-flag indicators. The total number of indicators marked as present or absent on the TPOT-P in the present study was 131. In addition, scores were calculated separately for environmental arrangements indicators ($v = 7$), key practices subcomponent indicators ($v = 108$), and red flags ($v = 16$).

CLASS. The CLASS (Pianta et al., 2008) is an observational, judgment-based rating scale designed to assess classroom quality focused on interactions and use of curricular materials in preschool to third-grade classrooms. It consists of 10 dimension items organized under three domains: (a) emotional support ($v = 4$), (b) classroom organization ($v = 3$), and (c) instructional support ($v = 3$). Scores for dimensions and domains on the CLASS range from 1 (*low*) to 7 (*high*). The instrument has been demonstrated to have interrater score agreement across the 10 dimensions ranging from 78.8% (regard for student perspectives) to 96.9% (productivity)

based on data collected in 164 preschool classrooms in Virginia (Pianta et al., 2008). Internal consistency score reliability estimates for the 10 CLASS dimensions range from .79 to .91 based on data collected during four observation cycles in 240 preschool classrooms. Criterion score validity estimates based on correlations with scores from the Early Childhood Environment Rating Scale, Revised Edition (ECERS-R; Harms, Clifford, & Cryer, 2005) range from .45 to .63 and from .23 to .43 for the Emerging Academics Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2001), a measure of the percentage of time spent in adult-elaborated interactions (Pianta et al., 2008).

Data Collection Procedures

TPOT-P. Data were collected in 50 preschool classrooms during a preschool year by two of six trained observers. Each teacher was observed on three occasions and each measurement occasion for each teacher was separated by 2 weeks. Each observer was provided a scoring manual and attended an 8-hr structured training to learn how to administer the TPOT-P. During training, observers viewed video illustrations to help them understand TPOT-P components, subcomponents, indicators, and scoring procedures. Before conducting observations for the study, each observer was required to have at least 80% interobserver agreement on total TPOT-P score with the trainer for three consecutive live observations. Observations in classrooms by the trained TPOT-P observers lasted at least 2 hr and included a mix of teacher-directed, child-initiated, and transition activities as well as a 10- to 15-min teacher interview conducted on the same day.

Before the first observation occasion, two raters were randomly selected from a pool of six trained raters to observe and score a TPOT-P in each classroom, such that each teacher was nested within a rater pair. The two raters assigned to observe and score the TPOT-P in a classroom on the first occasion observed and scored in the same classroom on subsequent measurement occasions so that the same rater pair observed a classroom on all three occasions. In total, there were 15 rater pairs and each rater pair was assigned to observe in one to six classrooms. Interobserver agreement was calculated for 100% of administrations for the total TPOT-P score and for the three TPOT-P components. Interobserver agreement was calculated using the following formula: ([smaller sum score / larger sum score] × 100). Average interobserver agreement for total TPOT-P score across all three rating occasions was 92% (*SD* = 9%). Average interobserver agreement for environmental arrangement, key practices subcomponents, and red flags were 93% (*SD* = 10%), 89% (*SD* = 9%), and 94% (*SD* = 7%), respectively.

CLASS. CLASS data collection occurred in all 50 classrooms between the second and third TPOT-P measurement occasions. Observers attended CLASS training conducted by certified trainers and were trained to CLASS interobserver agreement standards. At each administration, raters conducted four cycles of observations as recommended in the CLASS manual. Each cycle included 20 min of observation time and 10 min of scoring time, for a total administration time of 2 hr. We calculated interobserver agreement using the exact agreement method described in the CLASS manual for 33% of observations for each of the three domains of classroom quality measured by the CLASS. Average interobserver agreement for emotional support, classroom organization, and instructional support was 91% (*SD* = 9%), 90% (*SD* = 10%), and 86% (*SD* = 8%), respectively.

Data Analysis Procedures

TPOT-P and CLASS scores. We calculated means and standard deviations of TPOT-P scores at every rating occasion and mean scores of each CLASS domain. CLASS dimension scores were calculated by summing the scores of each dimension across the four observation cycles and dividing by 4 to obtain a mean score for each dimension. CLASS domain scores were calculated to represent the average of each of the corresponding dimension scores according to instructions provided in the CLASS manual (Pianta et al., 2008). We calculated bivariate correlations for teachers' total TPOT-P scores between each pair of measurement occasions. The purpose of these descriptive analyses was to examine the central tendency and variability in scores and, for the TPOT-P, to examine variability in scores within and across measurement occasions and participants.

Generalizability studies. A defining feature of the TPOT-P is its application within the context of preschool classrooms. Although use of contextualized instruments is necessary to capture authentically the behaviors of teachers and young children, observations conducted in naturalistic contexts such as classrooms often introduce potential sources of error variance in observed scores, which can impact score reliability or dependability (Bruckner, Yoder, & McWilliam, 2006). The classical test theory true score model articulated by Spearman (1907; 1913) is based on the premise that any observed score is the composite of a hypothetical true score for an individual and error (Crocker & Algina, 2008). Using classical test theory, a researcher can estimate one source of error variance at a time. For example, variation in observed scores associated with different raters can be assessed with interrater score reliability, or variation in observed scores on different occasions can be assessed with test−retest score reliability (Haertel, 2006; Shavelson & Webb, 1991; Thompson, 2003).

Generalizability (G) theory extends classical test theory and concepts about score reliability as a method for analyzing dependability of measurement by taking into account multiple sources of error variance simultaneously (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Thompson, 2003). Generalizability studies can yield two reliability coefficients corresponding to two different decisions associated with score inferences drawn from interpretation of observed scores. The *G* coefficient provides information about the consistency of observed scores and is used to determine how the object of measurement (e.g., teachers) performs relative to others, whereas the *Phi* coefficient provides information about the consistency of observed scores compared with a specific criterion. Information from G studies can be used to conduct decision (D) studies, in which researchers investigate the effects of multiple measurement conditions (e.g., number of administrations, number of raters) on the measurement dependability of observed scores (Shavelson & Webb, 1991). D studies allow researchers to make decisions to maximize resources while maintaining the integrity of measurement processes to produce dependable scores. G studies are optimal analytic procedures for examining the dependability of scores for measures designed for use in authentic early childhood settings, because they allow researchers to investigate simultaneously multiple facets of the measurement process (e.g., raters, occasions, items) associated with variation in observed scores (Bruckner et al., 2006; Goodwin & Goodwin, 1991; McWilliam & Ware, 1994).

In the present study, we conducted G studies to estimate the variance in observed TPOT-P scores associated with classroom teachers (i.e., the object of measurement), raters, rating occasions, and TPOT-P indicators. We estimated variance components for four sets of fidelity indicators (i.e., all TPOT-P indicators and indicators associated with each of the three TPOT-P components). We then conducted D studies to help determine the optimal conditions for maximizing the dependability of TPOT-P scores and using limited resources most efficiently in future feasibility and efficacy studies.

G-study analyses. Our G-study design most closely represented a crossed, three-facet model. In addition to the object of measurement (i.e., teachers), we posited that raters, occasions, and items were potential sources of error (facets) affecting the dependability of observed TPOT-P scores. The design was T × R × I × O; however, it was not possible for every rater to rate every teacher. Instead, a pair of raters rated each teacher. There were 15 rater pairs. The sample size for rater pairs ranged from one to six teachers. Three rater pairs had a sample size of one teacher; 2 rater pairs had a sample size of two teachers; 3 rater pairs had a sample size of three teachers; 5 rater pairs had a sample size of four teachers; 2 rater pairs had a sample size of five teachers; and 1 rater pair had a sample size of six teachers. We used restricted maximum likelihood implemented in PROC HPMIXED in SAS 9.2 to estimate the variance components. The HPMIXED procedure is designed to estimate variance components in large mixed-models by using sparse matrix techniques. Several of our models were large due to the number of levels in the item facet.

We conducted four G-study analyses, each with different sets of fidelity indicators. The indicator sets of interest in the present study were (a) all indicators (v = 131), (b) environmental arrangement indicators (v = 7), (c) indicators associated with 14 of the 15 TPOT-P key practices subcomponents (v = 108), and (d) "red-flag" indicators (v = 16). We calculated *Phi* and *G* coefficients for each indicator set using the variance components estimated by the HPMIXED procedure. The *Phi* coefficient was calculated using the following formula:

$$Phi = \frac{\sigma^2(t)}{\begin{aligned}&\sigma^2(t)+\sigma^2(r)/\mathrm{n'}(r)+\sigma^2(i)/\mathrm{n'}(i)+\sigma^2(o)/\mathrm{n'}(o)+\sigma^2(tr)/\mathrm{n'}(r)+\\&\sigma^2(ti)/\mathrm{n'}(i)+\sigma^2(to)/\mathrm{n'}(o)+\sigma^2(ri)/\mathrm{n'}(r)\mathrm{n'}(i)+\sigma^2(ro)/\mathrm{n'}(r)\mathrm{n'}(o)+\\&\sigma^2(io)/(\mathrm{n'}(i)\mathrm{n'}(o)+\sigma^2(tro)/(\mathrm{n'}(r)\mathrm{n'}(o)+\sigma^2(tri)/(\mathrm{n'}(r)\mathrm{n'}(i)+\\&\sigma^2(tio)/(\mathrm{n'}(i)\mathrm{n'}(o)+\sigma^2(rio)/(\mathrm{n'}(r)\mathrm{n'}(i)\mathrm{n'}(o)+\sigma^2(trio,e)/(\mathrm{n'}(r)\mathrm{n'}(i)\mathrm{n'}(o)\end{aligned}}.$$

The *G* coefficient was calculated using the following formula:

$$G = \frac{\sigma^2(t)}{\begin{aligned}&\sigma^2(t)+\sigma^2(r)/\mathrm{n'}(r)+\sigma^2(tr)/\mathrm{n'}(r)+\sigma^2(ti)/\mathrm{n'}(i)+\sigma^2(to)/\mathrm{n'}(o)+\\&\sigma^2(ri)/\mathrm{n'}(r)\mathrm{n'}(i)+\sigma^2(ro)/\mathrm{n'}(r)\mathrm{n'}(o)+\sigma^2(tro)/(\mathrm{n'}(r)\mathrm{n'}(o)+\\&\sigma^2(tri)/(\mathrm{n'}(r)\mathrm{n'}(i)+\sigma^2(tio)/(\mathrm{n'}(i)\mathrm{n'}(o)+\sigma^2(rio)/(\mathrm{n'}(r)\mathrm{n'}(i)\mathrm{n'}(o)+\\&\sigma^2(trio,\,e)/(\mathrm{n'}(r)\mathrm{n'}(i)\mathrm{n'}(o)\end{aligned}}.$$

Each equation shows how a generalizability coefficient is calculated from the variance components for teachers (t), raters (*r*), items (*i*), and rating occasions (*o*), where each variance component except the object of measurement (*t*) is divided by the number of conditions for each source of error represented in the variance component (Shavelson & Webb, 1991). In calculating the coefficient for absolute decisions (i.e., *Phi*), all variance components are included, whereas the equation for calculating the coefficient for relative decisions (i.e., *G*) includes only variance components that might conceivably contribute to differences in the object of measurement (i.e., teachers). The equation used to calculate the *G* coefficient for the present study yielded conservative estimates of *G*, because it assumed a different rater for each classroom. In the present study, there were actually subsets of teachers who were rated by the same raters; therefore, actual *G* coefficients would be expected to be higher than those reported.

D-study analyses. After the G studies, we conducted D-study analyses on the four primary TPOT indicator sets of interest. The purpose of the D studies was to evaluate changes in *Phi* and *G* if the TPOT-P was used by one rater on one, two, three, five, or seven occasions.

Convergent validity analyses. In addition to yielding dependable scores, assessment of fidelity should also result in scores that provide an accurate representation of practice relative to the construct(s) of interest for a given interpretation or use. Validity refers to the degree to which one can justify particular inferences drawn or actions taken based on measurement scores (Messick, 1990). The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999) outlined five sources of evidence for validity: (a) content, (b) substantive, (c) structural, (d) external relationships, and (e) consequences. In the present article, we describe our preliminary investigations of validity based on convergent evidence (evidence for external relationships). We conducted convergent score validity studies of the TPOT-P with the CLASS for each of the 50 teachers/classrooms. We examined bivariate Pearson product−moment correlations between four TPOT-P summary scores (total indicators, environmental arrangements, red flags, key practices subcomponents) and the 10 dimension and three domain scores for the CLASS.

Results

TPOT-P and CLASS Scores

The mean total TPOT-P scores (v = 131) at each of the three measurement occasions were 65.1 (SD = 15.24), 61.2 (SD = 16.1), and 58.6 (SD = 16.1), respectively. Across the 50 classrooms, total scores ranged from 35 to 106, indicating teachers were implementing between 26.7% and 80.9% of the indicators included on the TPOT-P. The mean scores for environmental arrangements (v = 7) at each of the three measurement occasions were 6.0 (SD = 0.9), 5.6 (SD = 0.8), and 6.0 (SD = 0.9), respectively. The mean scores for the key practices subcomponent fidelity indicators (v = 108) at each of the three measurement

occasions were 45.7 (*SD* = 13.6), 42.3 (*SD* = 14.3), and 39.9 (*SD* = 14.1), respectively. The mean scores for reverse-scored red flags ($v = 16$) at each measurement occasion were 13.4 (*SD* = 2.2), 12.9 (*SD* = 2.7), and 12.8 (*SD* = 2.4), respectively. The Pearson product−moment correlation coefficient for total TPOT-P scores ($v = 131$) between the first and second measurement occasion was .92, between the second and third measurement occasion was .90, and between the first and third measurement occasion was .87.

The descriptive statistics show the relative stability of scores for each of the four indicator sets. As would be expected, there was slightly more variability across measurement occasions in sets with more indicators. For all but one of the fidelity indicator sets, variability across classrooms within each measurement occasion was greater than the variability in scores across measurement occasions.

Means and standard deviations of the total score for each CLASS domain (i.e., emotional support, classroom organization, instructional support) were calculated for the CLASS administration that occurred between the second and third TPOT-P administration. The mean total score for emotional support was 4.5 (*SD* = 1.4); the mean score for classroom organization was 4.1 (*SD* = 1.3); the mean score for instructional support was 2.3 (*SD* = 1.2).

G and D Studies

In the G studies, indicators were the largest source of error variance in the statistical model, whereas raters and occasions were the smallest sources of error variance (see Table 2 for variance component estimates and percentage of variance associated with model facets). The *Phi* and *G* coefficients when all fidelity indicators were included were high, at .89 and .94, respectively, averaging over the three occasions and six raters. The *Phi* and *G* coefficients for the key practices indicators were also high, at .89 and .95. The *Phi* and *G* coefficients for the red flag indicators were .76 and .84, respectively. The *Phi* and *G* coefficients were lower for the environmental arrangement indicators (*Phi* = .23; *G* = .29). Table 3 shows the *Phi* and *G* coefficients from the D studies for all four of these TPOT-P indicator sets using one rater on one, two, three, five, and seven measurement occasions. Three of the four indicator sets had moderate to large *Phi* and *G* coefficients; the exception was environmental arrangements.

Convergent Validity

Pearson product−moment correlations between total TPOT-P scores and CLASS domain scores were .64 for emotional support, .69 for classroom organization, and .74 for instructional support. For key practices subcomponents scores and CLASS domain scores, correlations were .70 for emotional support, .73 for classroom organization, and .76 for instructional support. For red flags and CLASS domains scores, correlations were −.70 for emotional support, −.64 for classroom organization, and −.55 for instructional support. Correlations between environmental arrangements scores and CLASS domain scores were .08 for emotional support, .13 for classroom organization, and .11 for instructional support.

Table 2
Variance Component Estimates and Percentage of Variance Associated with Model Facets

	All indicators		Environmental arrangements		Key practices indicators		Red flags	
Variance components	σ^2	% of σ^2	σ^2	% of σ^2	σ^2	% of σ^2	σ^2	% of σ^2
Teacher (*t*)	.013	5.1	.003	2.55	.015	6.12	.017	10.83
Rater (*r*)	0	0	0	0	0	0	0[a]	0
Occasion (*o*)	.001	0.22	0	0	.001	<1	0[a]	0
Indicator (*i*)	.091	36.42	.021	16.99	.072	30.00	.034	21.74
Rater × Teacher (*rt*)	0[a]	0	0[a]	0.035	0[a]	0	0	0
Occasion × Teacher (*ot*)	.001	0.37	0[a]	0.09	.001	0.39	.002	1.59
Indicator × Teacher (*it*)	.033	13.09	.045	0.36	.031	12.86	.027	17.63
Rater × Occasion (*ro*)	0	0	0	0	0	0	0	0
Rater × Indicators (*ri*)	.011	4.26	.004	3.34	.012	4.94	.005	3.39
Occasion × Indicator (*oi*)	0[a]	0	0[a]	0	0[a]	0	0[a]	0
Rater × Teacher × Occasion (*rto*)	0[a]	0	0	0	.001	<1	0	0
Rater × Teacher × Indicator (*rti*)	.001	3.48	.010	7.91	.010	4.02	.001	<1
Rater × Occasion × Indicator (*roi*)	0[a]	0	0	0	0[a]	0.10	0	0
Occasion × Teacher × Indicator (*oti*)	.030	11.15	.012	9.43	.030	12.57	.018	11.91
Error (*rtio, e*)	.064	25.42	.029	23.48	.068	28.21	.049	31.54

[a]Actual estimates were negative and close to zero. Zero values were substituted as recommended by Cronbach, Gleser, Nanda, and Rajaratnam (1972).

Table 3
***Phi* and *G* Coefficients From D Studies**

	All indicators (v = 131)		Environmental arrangement indicators (v = 7)		Fourteen key practices indicators (v = 108)		Red flag indicators (v = 16)	
No. of observations	*Phi*	*G*	*Phi*	*G*	*Phi*	*G*	*Phi*	*G*
1	.76	.82	.15	.18	.76	.82	.60	.65
2	.82	.88	.18	.22	.82	.87	.68	.75
3	.85	.90	.19	.23	.85	.90	.72	.81
5	.87	.92	.20	.25	.87	.91	.75	.83
7	.88	.93	.20	.25	.88	.92	.76	.84

Discussion

The TPOT-P was designed to measure the fidelity with which preschool teachers implement practices associated with the *Pyramid Model* framework. In this article, we presented a case example of how we developed an assessment instrument for measuring intervention

fidelity while simultaneously considering its intended future use as a measure of implementation fidelity. We illustrated the processes used to identify indicators for the fidelity instrument. Analytic procedures used to examine preliminary psychometric integrity evidence were described along with results from these analyses.

Findings from the descriptive analyses suggested that without professional development focused on the *Pyramid Model* and associated practices, preschool teachers were implementing, on average, about 50% of the 131 indicators on the TPOT-P and the average level of practice implementation remained relatively consistent across the three measurement occasions. Although 86% of the teachers indicated they had received training in the previous year on topics related to social-emotional development and challenging behavior, only 3 of the 50 teachers were implementing more than 70% of the indicators on the TPOT-P and only 1 teacher was implementing 81% of indicators. Measures of variability indicated differences in implementation of TPOT-P practices were greater within each measurement occasion (across classrooms) than across each measurement occasion.

The stability of teachers' implementation of TPOT-P practice indicators across measurement occasions showed consistent implementation and little change in the number or percentage of practices implemented. This finding suggests targeted professional development or implementation support might be needed to alter TPOT-P scores significantly over time. In addition, there were sufficient numbers of indicators that were not credited across each of the measurement occasions to suggest ceiling effects would likely not occur in the absence of the *Pyramid Model* intervention (e.g., baseline measurement in single-subject experimental studies, preintervention in a potential efficacy or efficacy study, in counterfactual condition during a potential efficacy or efficacy trial).

Findings from the G study showed dependable observed scores across multiple raters and multiple occasions in applications where teachers' implementation of practices is criterion-referenced (*Phi* coefficient) and in situations where teachers' implementation is measured relative to others (*G* coefficient) for three of the primary indicator sets of interest (i.e., all indicators, indicators associated TPOT key practices subcomponents, red flags). Given the small proportion of variance associated with raters and occasions, findings from the D study suggest fewer occasions could be used to obtain dependable estimates of teacher implementation of *Pyramid Model* practices on all TPOT-P indicators, on the indicators associated with the 14 key practices subcomponents, and on the red flags. We chose not to conduct D-study analyses for the environmental arrangements indicators given the findings from the G study.

Of the four primary indicator sets of interest, those with fewer indicators (i.e., environmental arrangements, red flags) had lower generalizability coefficients regardless of the number of observations in the D study, suggesting a need for either more indicators or more administrations. Although each of the 14 key practices subcomponents represents a distinct practice (e.g., teaching children to express emotions, schedules, and routines), when scores were analyzed at the subcomponent level with 4 and 10 indicators associated with each subcomponent, generalizability coefficients, as expected, were lower. Indicators organized under each of the 14 key practices subcomponents are used to inform decisions about the implementation of practices associated with the subcomponent rather than to yield a score for interpretation at the individual subcomponent level. Results from the present study informed the decision to focus on three indicator sets of interest when interpreting TPOT-P

scores in the subsequent potential efficacy study (i.e., all indicators, indicators associated with the 14 key practices subcomponents, and the red flag indicators) because these three sets had the largest generalizability coefficients.

In all analyses, the variance in observed scores associated with indicators and averaged across classrooms, raters, and occasions was relatively high, and was always higher than the variance associated with teachers (objects of measurement). There are two possible explanations for this pattern of results. First, the TPOT-P was intended to measure implementation of practices associated with the multicomponent *Pyramid Model*; therefore, heterogeneity of indicators should be substantial to dependably measure fidelity of teacher practices associated with the model. Although indicator heterogeneity is an important feature of this instrument, it leads to more variation in observed scores associated with indicators, necessitating a relatively large number of indicators to yield consistent scores over multiple administrations and with multiple raters. In addition, the variance in observed scores associated with each indicator set reflects the variability in teachers' implementation of the individual practices associated with each indicator set across measurement occasions. Large generalizability coefficients were obtained because the total scores for each indicator set remained relatively constant across observations. Variable implementation of some individual indicators would not be unexpected given the day-to-day fluctuations in practices that would be expected across different observation occasions in preschool classrooms.

Convergent score validity evidence between TPOT-P scores and CLASS domain scores showed noteworthy correlations for TPOT-P total scores, key practices subcomponents scores, and red flags. Although our initial hypothesis was that TPOT-P scores would correlate with the emotional and instructional support dimensions of the CLASS, we found convergent evidence for all three domains. Noteworthy associations between scores on these measures are reasonable because higher fidelity of implementation of TPOT-P indicators should be associated with higher classroom instructional and interactional quality as measured by the CLASS. In addition, lower numbers of red flags should be associated with higher instructional and interactional quality as measured by the CLASS. The magnitude of the correlations between environmental arrangements scores and CLASS domain scores were small, which was not unexpected given the CLASS does not directly measure environmental features. Taken together, these results generally supported our initial hypotheses with regard to convergent score validity between the TPOT and the CLASS.

Although results from the generalizability and convergent validity analyses yielded promising preliminary psychometric evidence for the TPOT-P, it is important to note several limitations. First, we acknowledge that the optimal study design for this investigation would have been a fully crossed design in which every teacher was rated by every rater on each of the three measurement occasions. Due to limited resources, raters were not fully crossed with teachers for the present study. To account for this issue, we chose to calculate *G* and *Phi* coefficients that were conservative with respect to our design. Second, in the present study, we did not examine sensitivity of TPOT-P scores to assess intervention fidelity in response to teachers' participation in a professional development intervention focused on *Pyramid Model* practices versus business-as-usual professional development. As part of a potential efficacy trial conducted subsequent to the present study, however, we examined the sensitivity of TPOT-P scores across intervention and counterfactual conditions with promising results (Fox, Hemmeter, & Snyder, in press).

Although assessment of fidelity in authentic early childhood settings presents potential challenges to score reliability and validity, results from the present study suggest it is feasible to develop, validate, and evaluate fidelity measures for use in these contexts. Measures such as the TPOT-P are important to advance research and practice in early childhood, particularly with respect to the measurement of intervention and implementation fidelity for multicomponent interventions such as the *Pyramid Model* (Snyder, McLaughlin, & Denney, 2011). Previous studies have used G theory to analyze measurement properties of observational instruments involving child behaviors (Bruckner et al., 2006; McWilliam & Ware, 1994). The present study demonstrated the use of G theory to examine the psychometric properties of a multimethod fidelity instrument associated with the *Pyramid Model*. Findings suggest the initial psychometric evidence for the TPOT-P is promising.

Based on findings from the present study and other studies conducted using the TPOT-P, revisions to the instrument have been made. Research is being conducted on the revised instrument, the Teaching Pyramid Observation Tool–Research Edition (Fox et al., in press). As additional research is conducted to examine which teaching and instructional practices support young children to acquire and master skills related to social competence and prevent challenging behavior, measures like the Teaching Pyramid Observation Tool will be useful for assessing intervention and implementation fidelity, while adhering to standards for educational and psychological measurement.

References

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

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*, 68-79.

Bond, G. R., Evans, L., Salyers, M. P., Williams, J., & Kim, H. W. (2000). Measurement of fidelity in psychiatric rehabilitation. *Mental Health Services Research*, *2*, 75-87.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (Eds.). (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: Committee on Early Childhood Pedagogy, Commission on Behavioral and Social Sciences and Education, National Academy Press, National Research Council.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Brown, W. H., Odom, S. L., & McConnell, S. R. (2008). *Social competence of young children: Risk, disability, and intervention*. Baltimore, MD: Brookes.

Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention*, *28*, 139-154. doi:10.1177/105381510602800205

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*, 166-176. doi:10.1016/j.ecresq.2009.10.004

Chien, N. C., Howes, C., Burchinal, M., Pianta, R. C., Ritchie, S., Bryant, D. M., . . .Barbarin, O. A. (2010). Children's classroom engagement and school readiness gains in pre-kindergarten. *Child Development*, *81*, 1534-1549.

Conroy, M. A., Brown, W. H., & Olive, M. L. (2008). Social competence interventions for young children with challenging behavior. In W. H. Brown, S. L. Odom, & S. R. McConnell (Eds.), *Social competence of young children: Risk, disability, and intervention* (pp. 205-232). Baltimore, MD: Brookes.

Conroy, M. A., Dunlap, G., Clarke, S., & Alter, P. J. (2005). A descriptive analysis of behavioral intervention research with young children with challenging behavior. *Topics in Early Childhood Special Education*, *25*, 157-166.

Crocker, C., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. R. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: John Wiley.

Cronbach, L. J., Rajaratnam, N. R., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *The British Journal of Statistical Psychology*, *16*, 137-163.

Denham, S. A., Blair, K. A., DeMulder, E., Levitas, J., Sawyer, K., Auerbach-Major, S., & Queenan, P. (2003). Preschool emotional competence: Pathway to social competence? *Child Development*, *74*, 238-256. doi:10.1111/1467-8624.00533

Domitrovich, C. E., Cortes, R., & Greenberg, M. T. (2007). Improving young children's social and emotional competence: A randomized trial of the Preschool PATHS Program. *Journal of Primary Prevention*, *28*, 67-91.

Downer, J. (2013). Applying lessons learned from evaluations of model early care and education programs to preparation for effective implementation at scale. In T. Halle, A. Metz, & I. M. Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 157-169). Baltimore, MD: Brookes.

Downer, J., & Yazejian, N. (2013, April). *Measuring the quality and quantity of implementation in early childhood interventions* (OPRE Research Brief OPRE 2013-12). Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Dunlap, G., Conroy, M., Kern, L., DuPaul, G., VanBrakle, J., Strain, P., . . .Joseph, G. E. (2003). *Research synthesis on effective intervention procedures: Executive summary*. Retrieved from http://www.challenging-behavior.org/explore/publications_docs/research_synthesis.pdf

Dunlap, G., Strain, P. S., Fox, L., Carta, J. J., Conroy, M., Smith, B. J., . . .Sowell, C. (2006). Prevention and intervention with young children's challenging behavior: Perspectives regarding current knowledge. *Behavioral Disorders*, *32*, 29-45.

Dunlap, G., Wilson, K., Strain, P., & Lee, J. K. (2013). *Prevent-teach-reinforce for young children: The early childhood model of individualized positive behavior support*. Baltimore, MD: Brookes.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (FMHI Publication No. 231). Tampa: University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Fox, L., & Hemmeter, M. L. (2009). A program-wide model for supporting social emotional development and addressing challenging behavior in early childhood settings. In W. Sailor, G. Dunlap, G. Sugai, & R. Horner (Eds.), *Handbook of positive behavior support* (pp. 177-202). New York, NY: Springer.

Fox, L., Carta, J., Strain, P. S., Dunlap, G., & Hemmeter, M. L. (2010). Response-to-intervention and the Pyramid model. *Infants & Young Children*, *23*, 3-14.

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*, 48-52.

Fox, L., Hemmeter, M. L., & Snyder, P. A. (in press). *Teaching pyramid observation tool–Research edition*. Baltimore, MD: Brookes.

Fox, L., Hemmeter, M. L., & Snyder, P. (2008, August). *Teaching Pyramid Observation Tool for preschool classrooms: Pilot version* (Unpublished instrument and manual). Nashville, TN: Vanderbilt University.

Fox, L., Hemmeter, M. L., Snyder, P. S., Binder, D. P., & Clarke, S. (2011). Coaching early childhood special educators to implement a comprehensive model for the promotion of young children's social competence. *Topics in Early Childhood Special Education*, *31*, 178-192.

Goodwin, L. D., & Goodwin, W. L. (1991). Using generalizability theory in early childhood special education. *Journal of Early Intervention*, *15*, 193-204. doi:10.1177/105381519101500208

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education.

Hagermoser Sanetti, L. M., Dobey, L. M., & Gritter, K. L. (2012). Treatment integrity of interventions with children in the Journal of Positive Behavior Interventions from 1999 to 2009. *Journal of Positive Behavior Interventions*, *14*, 29-46. doi:10.1177/109830071145853

Halle, T., Metz, A., & Martinez-Beck, I. (2013). *Applying implementation science in early childhood programs and systems*. Baltimore, MD: Brookes.

Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early Childhood Environment Rating Scale Revised Edition*. New York, NY: Teachers College Press.

Hemmeter, M. L., Fox, L., & Snyder, P. (2007). *Examining the potential efficacy of a classroom wide model for promoting social emotional development and addressing challenging behavior in preschool children with and without disabilities* [Abstract]. Retrieved from http://ies.ed.gov/funding/grantsearch/details.asp?ID=577

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. 283-300). Baltimore, MD: Brookes.

Hemmeter, M. L., Ostrosky, M., & Fox, L. (2006). Social and emotional foundations for early learning: A conceptual model for intervention. *School Psychology Review*, *35*, 583-601.

Hemmeter, M. L., Snyder, P., Fox, L., & Algina, J. (2011, April). *Efficacy of a classroom wide model for promoting social-emotional development and preventing challenging behavior*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

Hulleman, C. S., Rimm-Kaufman, S. E., & Abry, T. (2013). Innovative methodologies to explore implementation: Whole-part-whole-construct validity, measurement, and analytical issues for intervention fidelity assessment in education research. In T. Halle, A. Metz, & I. M. Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 65-93). Baltimore, MD: Brookes.

Joseph, G. E., & Strain, P. S. (2003). Comprehensive evidence-based social emotional curricula for young children: An analysis of efficacious adoption potential. *Topics in Early Childhood Special Education*, *23*, 65-76.

LaParo, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal*, *104*, 409-426.

LeLaurin, K., & Wolery, M. (1992). Research standards in early intervention: Defining, describing, and measuring the independent variable. *Journal of Early Intervention*, *16*, 275-287.

Lloyd, C. M., Supplee, L. H., & Mattera, S. K. (2013). An eye to efficient and effective fidelity measurement for both research and practice. In T. Halle, A. Metz, & I. M. Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 139-155). Baltimore, MD: Brookes.

Mashburn, A. J., & Pianta, R. (2010). Opportunity in early education: Improving teacher-child interactions and child outcomes. In A. Reynolds, A. Rolnick, M. Englund, & J. Temple (Eds.), *Childhood programs and practices in the first decade of life: A human capital integration* (pp. 243-265). New York, NY: Cambridge University Press.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., . . .Howes, C. (2008). Measures of pre-k quality and children's development of academic, language and social skills. *Child Development*, *79*, 732-749.

McClelland, M. M., Morrison, F. J., & Holmes, D. L. (2000). Children at risk for early academic problems: The role of learning-related social skills. *Early Childhood Research Quarterly*, *15*, 307-329.

McLaren, E. M., & Nelson, C. M. (2009). Using functional behavior assessment to develop behavior interventions for children in Head Start. *Journal of Positive Behavior Interventions*, *11*, 3-21.

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention*, *18*, 34-46. doi:10.1177/105381519401800104

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ: Educational Testing Service.

Metz, A., & Bartley, L. (2012). Active implementation frameworks for program success: How to use implementation science to improve outcomes for children. *Zero to Three*, *32*(4), 11-18.

Metz, A., Halle, T., Bartley, L., & Blasberg, A. (2013). The key components of successful implementation. In T. Halle, A. Metz, & I. M. Beck (Eds.), *Applying implementation science in early childhood programs and systems* (pp. 21-42). Baltimore, MD: Brookes.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315-340. doi:10.1177/109821400302400303

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*, 33-84. doi:10.3102/0034654307313793

Pianta, R. C., LaParo, K., & Hamre, B. (2008). *Classroom Assessment Scoring System–PreK [CLASS]*. Baltimore, MD: Brookes.

Raver, C., & Knitzer, J. (2002). *Ready to enter: What research tells policymakers about strategies to promote social and emotional school readiness among three- and four-year old children*. New York, NY: National Center for Children in Poverty.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2001). *Emerging academics snapshot* (Unpublished instrument), University of California, Los Angeles.

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

Shonkoff, J. P. (2010). Building a new biodevelopmental framework to guide the future of early childhood policy. *Child Development*, *81*, 357-367.

Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academy Press.

Snyder, P., McLaughlin, T., & Denney, M. (2011). Frameworks for guiding program focus and practices in early intervention. In J. M. Kauffman & D. P. Hallahan (Series Eds.) & M. Conroy (Section Ed.), *Handbook of special education: Section XII Early identification and intervention in exceptionality* (pp. 716-730). New York, NY: Routledge.

Spearman, C. (1907). Demonstration of formulae for the true measurement of correlation. *American Journal of Psychology*, *18*, 72-101.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology*, *5*, 417-426.

Strain, P. S., & Joseph, G. E. (2006). You got to have friends. *Young Exceptional Children Monograph Series*, *8*, 1-22.

Thompson, B. (2003). *Score reliability*: *Contemporary thinking on reliability issues*. Thousand Oaks, CA: SAGE.

Thompson, R. A., & Goodman, M. (2009). Development of self, relationships, and socioemotional competence: Foundations for early school success. In O. A. Barbarin & B. H. Wasik (Eds.), *The handbook of child development and early education: Research to practice* (pp. 147-171). New York, NY: Guilford.

Thompson, R. A., & Raikes, H. A. (2007). The social and emotional foundations of school readiness. In D. F. Perry, R. K. Kaufmann, & J. Knitzer (Eds.), *Social and emotional health in early childhood: Building bridges between services and systems* (pp. 13-36). Baltimore, MD: Brookes.

Vaughn, S., Kim, A., Sloan, C. V. M., Hughes, M. T., Elbaum, B., & Sridhar, D. (2003). Social skills interventions for young children with disabilities: A synthesis of group design studies. *Remedial and Special Education*, *24*, 2-15.

Webster-Stratton, C. (1999). *How to promote children's social and emotional competence*. Thousand Oaks, CA: SAGE.

Webster-Stratton, C., & Reid, M. J. (2003). Treating conduct problems and strengthening social and emotional competence in young children: The Dina Dinosaur treatment program. *Journal of Emotional and Behavioral Disorders*, *11*, 130-143.

Wolery, M. (2011). Intervention research: The importance of fidelity measurement. *Topics in Early Childhood Special Education*, *31*, 155-157. doi:10.1177/0271121411408621