Clinical Validity of a Brief Measure of Early Childhood Social–Emotional/Behavioral Problems

Margaret J. Briggs-Gowan,1 PhD, Alice S. Carter,2 PhD, Kimberley McCarthy,1 BA, Marilyn Augustyn,3 MD, Elizabeth Caronna,4 MD, and Roseanne Clark,5 PhD

1Department of Psychiatry, University of Connecticut Health Center, 2Department of Psychology, University of Massachusetts at Boston, 3Department of Pediatrics, Boston University, 4Harvard Vanguard Medical Associates, Developmental and Behavioral Pediatrics, Boston, MA, and 5Department of Psychiatry, University of Wisconsin School of Medicine and Public Health

All correspondence concerning this article should be addressed to Margaret J. Briggs-Gowan, PhD, Department of Psychiatry, University of Connecticut Health Center, 263 Farmington Avenue, MC-1410, Farmington, CT 06030, USA. E-mail: mbriggsgowan@uchc.edu

Received June 27, 2012; revisions received February 3, 2013; accepted February 16, 2013

Objective To address a pressing need for measures of clinically significant social–emotional/behavioral problems in young children by examining several validity indicators for a brief parent-report questionnaire. Methods An ethnically and economically diverse sample of 213 referred and nonreferred 2- and 3-year-olds was studied. The validity of the Brief Infant–Toddler Social–Emotional Assessment (BITSEA) Problem Index and Internalizing and Externalizing scales was evaluated relative to a “gold standard” diagnostic interview, as well as the Child Behavior Checklist (CBCL). Results The validity of the BITSEA Problem Index relative to Diagnosis (sensitivity = 72.7%–80.8%, specificity = 70.0%–83.3%) and clinical-range CBCL scores (sensitivity = 80.0%–96.2%, specificity = 75.0%–89.9%) was supported in the full sample and within minority/nonminority groups. Additional results supported the validity of the BITSEA Internalizing and Externalizing scales. Conclusions Documented validity suggests that the BITSEA may be a valuable tool to aid screening, identification, and assessment efforts targeting early-emergent social–emotional/behavioral problems. Practical implications and generalizability are discussed.

Key words assessment; early childhood; psychiatric diagnosis; screening; social–emotional/behavioral problems; surveillance.

Introduction

Recent decades have witnessed remarkable progress in early childhood mental health assessment and intervention, with increased recognition of the importance of this period to brain development and lifespan functioning (Teicher, Tomoda, & Andersen, 2006). Accumulating evidence documents the presence of clinically significant social–emotional/behavioral problems in children aged ≤5 years, with prevalence estimates for psychiatric disorders in this age group ranging from ~16%–26% (cf., Egger & Angold, 2006; Lavigne, Lebailly,Hopkins, Gouze, & Binns, 2009). Moreover, young children are known to suffer from both externalizing and internalizing problems, with prevalence estimated at 9% and 11%, respectively (Egger & Angold, 2006). Early problems persist and predict later mental health problems and poorer academic performance (e.g., Briggs-Gowan & Carter, 2008; Campbell, Spieker, Burchinal, Poe, & NICHD Early Child Care Research Network, 2006; Luby, Si, Belden, Tandon, & Spitznagel, 2009; Shaw, Lacourse, & Nagin, 2005). In the preschool period, disruptive behavior problems are among the most well-recognized and treated mental health problems (Brownell & Carriger, 1990), whereas internalizing problems are often overlooked, despite their prevalence, persistence and prediction of poorer outcomes later in childhood (Bayer et al., 2010; cf., Egger & Angold, 2006).
Recent strides in treatment have yielded evidence-based treatments for early-emergent difficulties in both domains (e.g., Bayer et al., 2010; Thomas & Zimmer-Gembeck, 2007). Yet, a long-standing history of failure to identify and treat early-childhood mental health problems remains (Costello, Egger, & Angold, 2005; Lavigne et al., 2009; Pavuluri, Luk, & McGee, 1996; Shedrick, Merchant, & Perrin, 2011; Stagman & Cooper, 2010). Highlighting this problem, in one recent study, only 3% of preschoolers with psychiatric disorders in a primary care sample had received services (Lavigne et al., 2009). This problem is particularly pronounced in minority children (Costello et al., 2005).

Parents and professionals alike have difficulty knowing “when to worry” about young children’s behavior. Many parents of young children with clinically significant problems are not worried and do not seek help (Ellingson, Briggs-Gowan, Carter, & Horwitz, 2004). There is also substantial variability in the effectiveness of pediatric primary care professionals in identifying early-emergent problems (Shedrick et al., 2011). Brief, norm-referenced, structured measures can aid professionals in identifying clinically concerning problems, as a first step in detection in which scores above a set threshold trigger additional follow-up (e.g., conversation with parent to assess concerns and further assessment). Such tools also are essential for state programs that require screening at well-child visits for all children receiving Medicaid under the Early Periodic Screening, Diagnosis, and Treatment Program (Health Resources and Services Administration, Maternal and Child Health Bureau, 2010), as well as for assessing service needs and treatment outcomes within home-visiting programs for at-risk families with young children (MDRC, 2011). Thus, need exists across multiple systems for reliable and valid tools that can support early identification and intervention for early-emergent social–emotional/behavioral problems. Consistent evidence of internalizing and externalizing problems in young children further underscores the need for tools addressing problems in both domains.

Several brief measures of social–emotional/behavioral problems exist for children aged ≤5 years (Briggs-Gowan, Godoy, & Carter, 2011). Two of these, the Ages and Stages Social–Emotional (Squires, Bricker, & Twombly, 2003) and the Brief Infant–Toddler Social and Emotional Assessment (BITSEA; Briggs-Gowan & Carter, 2006), are particularly promising because they are designed for children as young as 1 year and have shown acceptable sensitivity and specificity relative to a validating criterion. However, to our knowledge, neither has established validity relative to concurrent psychiatric diagnosis nor sensitivity to internalizing and externalizing disorders more specifically. The focus of this study is the BITSEA, a measure that has demonstrated acceptable global validity relative to “any” Child Behavior Checklist (CBCL) problem (Briggs-Gowan, Carter, Irwin, Wachtel, & Cicchetti, 2004). However, validity in assessing psychiatric disorders and more specific internalizing and externalizing problems has not been examined.

In this study, the validity of the BITSEA is evaluated against the “gold standard” of psychiatric diagnosis established with a standardized interview, the Preschool Age Psychiatric Assessment (PAPA) (Egger et al., 2006), and against clinical status on the CBCL (Achenbach & Rescorla, 2000), a widely adopted dimensional checklist. This approach provides a test of the BITSEA’s ability to identify children whose problems meet the severity, duration, onset, and impairment demands of diagnostic criteria (children who are thus crucial to identify), as well as those with elevated symptoms, consonant with the movement toward dimensional approaches to psychopathology (Insel et al., 2010).

The over-arching goal of this study is to evaluate the validity and clinical utility of the BITSEA. The first aim is to examine the validity of the BITSEA Problem, Internalizing and Externalizing continuous scores through correlations with parallel symptom scores from the PAPA. The second aim is to evaluate the sensitivity and specificity of the BITSEA Problem Index in relation to “any” disorder by examining the validity of two BITSEA Problem cutscores, a more stringent Clinical cutscore and an Of Concern score that uses a lower threshold to detect less severe scores that may indicate risk for emergent psychopathology. Within this aim, attention is paid to the potential impacts of ethnicity, sociodemographic risk, and developmental status on indicators of validity. The third aim is to evaluate the sensitivity of the two BITSEA Problem Of Concern and Clinical cutscores and domain-specific Internalizing and Externalizing cutscores relative to internalizing and externalizing disorders. The fourth aim is to explore the clinical utility of the BITSEA domain-specific scores, by examining these scores across the following groups: no diagnosis, internalizing diagnosis, externalizing diagnosis, and both internalizing and externalizing diagnoses. The final aim is to evaluate the sensitivity and specificity of the BITSEA Problem Index cutscores in relation to the presence of elevated symptoms on the CBCL.

**Method**

**Participants**

The sample was composed of 213 children, aged 24–48 months, living with a custodial parent. Referred children
(n = 103) were recruited from six sites providing mental health and/or developmental assessments and services. Two referred sites provided mental health services for high-risk mothers and children (n = 43); three were Part C-funded early intervention programs that provided developmental assessments and services (n = 30), and one site was a developmental/behavioral pediatric clinic that provided assessments for a broad range of developmental and behavioral concerns (n = 30). Nonreferred children (n = 110) were recruited from the same communities as the referral sites. Three hundred ten parents consented and 260 (83.9%) completed the study protocol. Ten children were excluded owing to incomplete data (n = 6) or global cognitive delays or suspected/ diagnosed Autism Spectrum disorders (n = 4). Thirty-seven children aged <24 months were excluded from analyses because they were too young for the diagnostic interview.

Children were a mean of 34.9 months (standard deviation [Std] = 6.8). Most children were boys (71.8%). The sample was ethnically diverse (45.5% White, 27.2% African American/Black, 15.0% Hispanic/Latino, 6.6% multi-ethnic minority, 5.7% other). Respondents were biological mothers (86.4%), biological fathers (4.2%), adoptive/foster parents (6.2%), and relatives (3.2%). Mean parental age was 33.8 years (Std = 7.4). Respondent education varied (9.6% less than high school, 20.6% high school degree/Graduate Equivalency Diploma, 27.3% some college, 42.6% college degree or greater). Most children came from two-parent homes (71.7%). Forty-five percent of families were of low-income group. Children’s average developmental quotient was 93.7 (Std = 20.2). Recruitment sources were similar in ethnicity, low-income status, single- versus two-parent home, and respondent education, but children from mental health sites tended to be slightly older than those from other sources (p < .05).

**Procedures**

Clinic staff introduced the study to referred families. Non referred children were recruited in pediatric clinics by research assistants and via posted fliers. Informed consent was obtained by clinic staff or study personnel. Parents completed questionnaires and a semi-structured interview. Research assistants administered a standardized developmental assessment to children. Ethical guidelines for mandated reporting of maltreatment were followed. Parents received $50–$80; compensation varied because some measures were part of intake procedures at some sites. All procedures were approved by four Institutional Review Boards.

**Measures**

**BITSEA**

The BITSEA is a developmentally appropriate measure of social–emotional/behavioral problems in 1- to 3-year-olds (Briggs-Gowan & Carter, 2003, 2006). The 31-item general Problem Index includes internalizing, externalizing, and dysregulation problems. Acceptable test-retest (r = .92) and inter-rater reliability (r = .74) have been documented previously (Briggs-Gowan & Carter, 2006). Internal consistency was 0.88 in this sample. The Problem Index yields Clinical (extreme 15% of a normative sample) and Of Concern (extreme 25% of a normative sample) cutscores. These scores were defined in a national standardization sample for 12- to 36-month-olds and an epidemiologic sample for 37- to 48-month-olds. The Internalizing scale has 11 depression, anxiety, and negative emotionality items (e.g., Seems nervous, tense or fearful; Cries or hangs onto you when you try to leave; and Cries or tantrums until exhausted). The Externalizing scale has 7 aggression, defiance, and activity/impulsivity items (Hits, shoves, bites or kicks other children; Restless and can’t sit still; and Destructive, breaks or ruins things on purpose). Internal consistency was acceptable (α = .80 and .82, respectively). Secondary data analysis indicated acceptable test-retest reliability (.85 and .80, respectively, n = 119) in a previously studied sample (Briggs-Gowan et al., 2004).

**PAPA**

This semi-structured interview provides developmentally sensitive assessment of Diagnostic and Statistical Manual-IV disorders with impairment in 2–5-year-olds (Egger et al., 2006), through a glossary-driven interview in which symptoms are established through interviewer appraisal of descriptions of children’s behavior, rather than strict parental endorsements of symptoms. Acceptable test-retest reliability has been reported (Egger et al., 2006). Internalizing Disorders (i.e., Generalized Anxiety Disorder, Separation Anxiety, Specific Phobia, Major Depression, Dysthymia, and Depression Not Otherwise Specified) and Externalizing Disorders (Oppositional Defiant Disorder, Conduct Problems, and Attention Deficit Hyperactivity Disorder [ADHD]) were assessed. Diagnostic scoring algorithms require that symptoms occur at a frequency greater than what is typical for preschoolers. Analyses used diagnostic variables (e.g., “any” externalizing disorder) and symptom sums calculated as total criterion symptoms across disorders in a domain. Teams were trained by the developers of the PAPA. The PAPA was administered by clinical psychologists, postdoctoral fellows, graduate students, and a research assistant with several years of
experience administering semi-structured psychiatric interviews.

CBCL/1.5-5

This 99-item parent-report checklist for children aged 1½–5 years has acceptable test-retest reliability ($r = .68–.92$) and validity (Achenbach & Rescorla, 2000). Internal consistency was acceptable ($\alpha = .91–.94$). Clinical-range scores ($T$-scores $\geq 63$) on the internalizing, externalizing and total scores were used.

Mullen Scales of Early Learning

Developmental level was assessed with the composite score of the Mullen, a norm-referenced, standardized administered assessment of children’s cognitive, language, and motor development (Mullen, 1995).

Sociodemographic Characteristics

Parents reported on basic sociodemographic factors. Low income was defined as income at or below the poverty level following federal Consumer Price Index guidelines based on annual household income and number of adults and children in the home. Receipt of income-based public assistance was used as a proxy for low income when key income or household data were missing.

Analytic Approach

Initial analyses (Aim 1) used Pearson correlations to examine the relations among continuous scores. Sensitivity-specificity analyses were used for three aims (2, 3, and 5). These analyses yield statistics that provide insight into a tool’s validity with respect to identification of children who may have clinically significant problems. With psychiatric disorder as the validating measure and the BITSEA as the tool being validated, sensitivity refers to the percentage of children with disorders flagged as positive on the BITSEA, $A/(A+C)$ (Figure 1). Specificity is the percentage of children without disorders who are negative on the BITSEA, $D/(B+D)$. Positive Predictive Value (PPV) is the percentage of children flagged as positive by the BITSEA who have disorders, $A/(A+B)$. PPV and prevalence are positively associated, such that PPV decreases as prevalence decreases, assuming sensitivity and specificity are constant. Negative Predictive Value (NPV) is the percentage of children without disorder that are negative on the BITSEA, $D/(C+D)$. For domain-specific analyses, specificity and PPV are not discussed because they have limited relevance when a multi-domain measure is compared with a single-domain validator (i.e., true problems in a domain not assessed by the validator will pull specificity and PPV down). It is recommended that tools used for early identification of developmental or social–emotional/behavioral problems should have minimum sensitivity and specificity of 70% and false positive rates of no greater than 30% of all screened (Cicchetti, Volkmar, Klin, & Showalter, 1995; Council on Children with Disabilities, 2001). We are unaware of a recommended minimum PPV for brief measures of child developmental or social–emotional problems. Low PPV is tolerated when failure to detect a problem is associated with an extremely poor outcome (e.g., death). For example, a minimum PPV of 10% is proposed for ovarian cancer screening (Hensley & Spriggs, 2004). In the area of mental health screening, Milgrom et al. (Milgrom, Mendelsohn, & Gemmill, 2011) have cogently argued that a screening tool can be considered effective if it successfully identifies a sub-population in which the prevalence of the problem is more common than in the original population. Also under Aim 2, logistic regression models were used to evaluate the potential impact of sociodemographic and developmental factors on validity, by comparing models with and without these factors controlled. Following guidelines for identifying confounding variables, a 10% or greater change in the parameter estimate for the BITSEA between the raw model and the parallel adjusted model was used to signify impact of these factors on validity. Finally (Aim 4), the clinical utility of the BITSEA Internalizing and Externalizing scores were evaluated with Analysis of Variance across diagnostic groups (No disorder, Internalizing only, Externalizing only, and Both domains).

Results

Forty-eight children (22.5%) met criteria for one or more disorder. These disorders were distributed across the internalizing ($n = 36$, 16.9%) and externalizing ($n = 27$, 12.7%) domains, with 15 (7.0%) children reported to have disorders in both domains. Rates of individual disorders were as follows: Separation Anxiety (4.7%), Generalized Anxiety (5.6%), Depression/Dysthymia (10.8%), ADHD (6.1%), and Disruptive Behavior Disorders (10.8%).

The CBCL revealed clinical scores in the internalizing ($n = 33$, 16.4%) and externalizing ($n = 43$, 20.9%)
domains, with substantial overlap across domains (n = 20, 10.0%).

Children with disorders came from all recruitment sources (29 mental health, 60.4%; 11 developmental, 22.9%, and 8 nonreferred, 16.7%). Forty-four percent of these children were 2-year-olds and 56% were 3-year-olds.

Correlations Between BITSEA Scores and PAPA Symptoms (Aim 1)

Moderate to strong correlations between BITSEA summary scores and PAPA symptom sums measuring the same constructs ($r = .57-.73$) supported convergent validity (Table I). Divergent validity was supported for BITSEA Externalizing, as its correlation with PAPA Externalizing was stronger than its correlation with PAPA Internalizing (Fischer’s $r$ to $z'$ transformation = 4.47, $p < .001$), but not for BITSEA Internalizing, which had similar correlations with both PAPA domains ($z' = 1.28, n.s.$).

Correspondence Between BITSEA Problem Index and Any Disorder (Aim 2)

Comparison of the two BITSEA Problem index scores with the presence of Any PAPA Disorder revealed sensitivity and specificity that were consistently above recommended thresholds (see Table II). For example, the Clinical index had sensitivity of 72.9%, specificity of 81.8%, and PPV of 53.8%, while the Of Concern index had sensitivity of 79.2%, specificity of 70.9%, and PPV of 44.2% in relation to the category of children with a diagnosed psychiatric disorder. Comparable validity results were observed when these associations were examined within minority and nonminority groups separately. Furthermore, results of two multivariate logistic regression models that examined whether child age, sex, developmental level, minority ethnicity, and low-income status affected the relationships between the Of Concern and Clinical index scores and Any Disorder indicated that these factors had minimal effects on these associations; the changes in logistic regression coefficients ($B$) between these index scores and diagnostic status with and without these factors in models were well below 10% (data not shown).

Table I. Correlations Between BITSEA Scales and PAPA Symptoms

<table>
<thead>
<tr>
<th>BITSEA scales</th>
<th>PAPA symptoms</th>
<th>Total</th>
<th>Internalizing</th>
<th>Externalizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total problems</td>
<td>0.68</td>
<td>0.55</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Internalizing</td>
<td>0.56</td>
<td>0.57</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>Externalizing</td>
<td>0.72</td>
<td>0.46</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>

Note: BITSEA = Brief Infant–Toddler Social–Emotional Assessment; PAPA = Preschool Age Psychiatric Assessment.

Sensitivity of BITSEA Cutscores to Internalizing and Externalizing Disorders (Aim 3)

The Problem Index Clinical and Of Concern index scores had acceptable sensitivity to internalizing (77.8% and 83.3%) and externalizing (77.8% and 85.2%) disorders (see Table I). For individual disorders, including anxiety disorders, depressive disorders, ADHD, and ODD/CD, sensitivity was acceptable (76.2%–95.7%, data not shown) (specificity is of limited relevance to this comparison and therefore is not discussed).

Examination of the BITSEA domain-specific scores relative to disorders in the same domains indicated sensitivity of 66.7% (unacceptable) and specificity of 82.9% for BITSEA Internalizing versus internalizing disorders (bottom of Table I). Sensitivity and specificity were acceptable (70.4% and 84.4%, respectively) for BITSEA Externalizing versus externalizing disorders.

Clinical Utility of BITSEA Domain Scores (Aim 4)

Analyses exploring the clinical utility of the BITSEA Internalizing and Externalizing scores examined whether they showed domain-specific elevations depending on the presence of internalizing and externalizing disorders. Four groups were generated: NONE (no disorder), INT (internalizing disorder only), EXT (externalizing disorder only), and BOTH (internalizing and externalizing disorders). BITSEA scores were converted to $z$-scores such that scores $\geq 1.5$ are interpreted as clinically significant relative to normative data (see Table III).

The internalizing analysis of variance revealed a significant effect of diagnostic group on BITSEA Internalizing scores, with large effect size. Compared with the NONE group, BITSEA Internalizing scores were significantly higher in the BOTH and INT groups ($p < .05$). The externalizing model revealed a significant effect of diagnostic group on BITSEA Externalizing scores, with large effect size. Compared with the NONE group, BITSEA Externalizing scores were significantly higher in the BOTH and EXT groups ($p < .05$). Unexpectedly, Externalizing scores were significantly higher in the INT versus NONE group. An ad hoc analysis explored whether these elevations might reflect the presence of symptoms that fell short of diagnostic criteria by examining PAPA Externalizing symptoms across diagnostic groups. As with the BITSEA, PAPA Externalizing symptoms were significantly higher in the INT group (Mean [MN] = 2.8, Std = 2.1) than in the NONE group (MN = 1.1, Std = 1.3, $t = 3.54, p < .01$). Differential patterns of BITSEA domain scores according to the presence of internalizing or externalizing disorders
support the clinical utility of using these more refined scores, rather than relying on the Total Problem Score alone for these purposes.

**Correspondence Between the BITSEA Cutscores and CBCL Status (Aim 5)**

When clinical-range scores on the CBCL (T ≥ 63) were used as the validating criterion, the BITSEA Problem Of Concern and Clinical cutscores had acceptable sensitivity (80.0%–96.2%) and specificity (75.0%–89.9%) in the full sample, as well as within minority and nonminority groups (Table IV). PPV ranged from 57.4% to 76.7%. Sensitivity and specificity also were acceptable for all four BITSEA cutscores when compared with elevated Internalizing and Externalizing scores on the CBCL (sensitivity: 79.1%–97.0%; specificity: 70.8%–92.0%; PPV: 39.5%–72.3%).

**Discussion**

Findings from this clinically and socioeconomically diverse sample support the validity of the BITSEA, a brief, parent-report measure designed to support early detection of social–emotional/behavioral problems in young children, in relation to concurrent psychiatric disorder. Validity was supported both for problems overall and within the internalizing and externalizing domains. Of relevance to the problem of unmet need for mental health services among children from minority backgrounds (Costello et al., 2005), validity was comparable across minority and nonminority children. That the BITSEA showed acceptable validity relative to the CBCL is positive from a practical perspective because it is substantially shorter than the CBCL, but not surprising given the similarity of the assessment approach for the two measures and prior evidence of good correspondence when comparing the BITSEA problem domain to “any” problem on the CBCL (Briggs-Gowan et al., 2004). This is the first demonstration that the BITSEA shows strong validity relative to psychiatric disorder (identified through semi-structured interview) and with respect to both internalizing and externalizing disorders. Specific findings, including validity relative to psychiatric diagnosis, implications for early detection and clinical utility, insights gained regarding clinical heterogeneity, and limitations are discussed.

**Implications for Identification and Assessment**

The BITSEA Problem Index encompasses a broad array of problems across the internalizing and externalizing domains. When compared with “any” psychiatric disorder, both the BITSEA Problem Index cutpoints (Clinical and Of Concern), respectively, showed acceptable sensitivity (72.9% and 79.2%) and specificity (81.8% and 70.9%). These patterns were further supported within nonminority
and minority children and negligibly affected by sociodemographic risk and developmental level. Further investigation focusing within broad domains revealed that both cutscores had acceptable sensitivity to internalizing and externalizing disorders (77.8%–85.2%) (specificity is not discussed because it is not relevant to these analyses). Together, these findings support the BITSEA’s potential to aid in the detection of a broad range of clinically impairing problems targeted by early screening and home visiting programs and for which evidence-based interventions now exist. It is important to emphasize that the BITSEA is not intended to be used for diagnostic purposes, but rather as an aid to help professionals identify children who may be experiencing clinically significant social–emotional/behavioral problems and hence benefit from additional assessment and follow-up to ascertain management, referral, or treatment needs.

Results provide practical guidance about which of the two Problem Index cutscores is appropriate depending on goals and setting. The Of Concern score offers the highest sensitivity to overall problems, as well as problems within the internalizing and externalizing domains. Thus, the greatest proportion of children with psychiatric disorders will be flagged by this score. However, there is a tradeoff because specificity tends to be lower (though still acceptable) than for the Clinical score. Therefore, more children without psychiatric disorders will be flagged by the Of Concern score than the Clinical score. So, the Of Concern score is most appropriate when optimizing sensitivity is important and follow-up resources to address problems that range in severity are available, such as within medical home or home-visiting programs in which assessment occurs in the context of ongoing targeted prevention or intervention. Indeed, when prevention is the focus, flagging children whose problems have not yet crossed the diagnostic threshold may even be desirable. In contrast, when the goal is secondary intervention or when resources are more constrained, the Clinical score is more appropriate because it has higher specificity and positive predictive value. Thus, relative to the Of concern score, a greater proportion of children identified by the Clinical score will have psychiatric disorders, often an important consideration when resources for additional follow-up and intervention are scarce.

More specifically, 53.8% of the children flagged by the Clinical score had disorders. This was ~2.4 times the rate of 22.5% in sample overall, indicating acceptable PPV for early identification efforts according to recommendations by Milgrom et al. (2011). Nonetheless, whichever cutscore is used, careful training is essential to ensure that results are interpreted appropriately and discussions with parents include an explanation that results are based on a brief tool that does not provide diagnostic information and reflects the parents’ appraisal of their child’s behaviors. In addition, when appropriate, feedback should be integrated with prompt follow-up to assess clinical significance and/or need for additional management, referral, or treatment.

Framing the observed validity within prior research is challenging because few studies have reported the sensitivity and specificity of brief measures against psychiatric diagnosis at any age. One study compared the Pediatric Symptom Checklist (PSC), a brief screener for children aged ≥4 years (Jellinek, Murphy, & Burns, 1986), and the CBCL with psychiatric diagnosis (Gardner, Lucas, Kolko, & Campo, 2007). Three cutpoints on the PSC were examined, revealing sensitivity from 42% to 57%, specificity from 78% to 86%, and PPV from 10% to 35%. The CBCL had sensitivity of 31%, specificity of 96%, and PPV of 58%. Thus, the results for the BITSEA Clinical threshold (sensitivity: 72.9%; specificity: 81.8%; PPV: 53.8%) were at least comparable with available estimates for similar tools used with older children.

### Domain-Specific BITSEA Scores

At a time when many researchers and professionals crave ever-shorted and more focused assessments, it is noteworthy that although their validity was generally supported, the BITSEA Internalizing and Externalizing scores were not as sensitive to disorders (especially internalizing) as the Problem Index. Specifically, correlations between the BITSEA domain-specific scores and parallel symptom scores from the diagnostic interview were moderate to strong, providing some support for validity. Additionally,
both BITSEA Internalizing and Externalizing had acceptable sensitivity and specificity relative to elevated scores in the same domains of the CBCL. When compared with psychiatric disorders, however, although the BITSEA Externalizing score had acceptable sensitivity (70.4%) and specificity (84.4%) to externalizing disorders, the BITSEA Internalizing score had insufficient sensitivity to internalizing disorders (sensitivity 66.7%, specificity 82.9%). The more general BITSEA Problem Index is therefore recommended over the BITSEA Internalizing score for identifying potential internalizing problems.

Other findings supported the clinical utility of the domain-specific scores. Specifically, BITSEA Internalizing scores were substantially elevated among children with internalizing disorders, with scores tending to fall in the clinical range (1½–2 standard deviations above the norm). These elevations differentiated children with internalizing disorders from those without disorders. The same pattern of findings was observed for BITSEA Externalizing scores, with substantial elevations among children with Externalizing disorders compared with those without disorders.

Yet, findings also revealed notable heterogeneity in young children’s symptom presentations, with elevations in one domain often accompanied by elevations in another. In addition to 31.3% of children with disorders evidencing both externalizing and internalizing disorders, the most obvious example of cross-domain presentation was that, according to both the BITSEA and PAPA, externalizing symptoms were significantly elevated in children with only internalizing disorders compared with those without disorders. Cross-domain correlations between the BITSEA and PAPA also were moderate. These patterns complement epidemiologic findings indicating that co-occurring cross-domain problems affected ~6% of young children and strongly predicted persistent problems over time (Briggs-Gowan, Carter, Bosson-Heenan, Guyer, & Horwitz, 2006). Together, these patterns highlight the importance of comprehensive assessment across domains, even when the presenting concerns may focus on a single domain.

Generalizability of Study Findings

Children were recruited from referred and non referred sources to obtain a sufficient number of children with disorders to test study aims. This design may influence generalizability because the prevalence of a problem affects PPV, with PPV decreasing as prevalence decreases. Approximately 22.5% of children in this sample had disorders, which is on the high end of prevalence estimates for young children of 16%–21% in non referred normative samples and 26% in a low-income sample (cf., Egger & Angold, 2006). Thus, PPV is expected to be somewhat lower than observed here in populations in which

<table>
<thead>
<tr>
<th>Measures evaluated</th>
<th>BITSEA negative/ CBCL negative</th>
<th>BITSEA positive/ CBCL negative</th>
<th>BITSEA negative/ CBCL positive</th>
<th>BITSEA positive/ CBCL positive</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>PPV (%)</th>
<th>NPV (%)</th>
<th>False positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>BITSEA Problem Index Cutscores versus Any elevated CBCL Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Full sample</td>
<td>133</td>
<td>16</td>
<td>9</td>
<td>47</td>
<td>83.9</td>
<td>89.3</td>
<td>74.6</td>
<td>93.7</td>
</tr>
<tr>
<td></td>
<td>Nonminority</td>
<td>62</td>
<td>7</td>
<td>3</td>
<td>23</td>
<td>88.5</td>
<td>89.9</td>
<td>76.7</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>71</td>
<td>9</td>
<td>6</td>
<td>24</td>
<td>80.0</td>
<td>88.8</td>
<td>72.7</td>
<td>92.2</td>
</tr>
<tr>
<td>Of concern</td>
<td>Full sample</td>
<td>117</td>
<td>32</td>
<td>4</td>
<td>52</td>
<td>92.9</td>
<td>78.5</td>
<td>61.9</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td>Nonminority</td>
<td>57</td>
<td>12</td>
<td>1</td>
<td>23</td>
<td>96.2</td>
<td>82.6</td>
<td>67.6</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td>Minority</td>
<td>60</td>
<td>20</td>
<td>3</td>
<td>27</td>
<td>90.0</td>
<td>75.0</td>
<td>57.4</td>
<td>95.2</td>
</tr>
<tr>
<td>BITSEA Problem Index Cutscores versus Elevated CBCL internalizing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Full sample</td>
<td>137</td>
<td>31</td>
<td>4</td>
<td>29</td>
<td>87.9</td>
<td>81.5</td>
<td>48.3</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>Nonminority</td>
<td>119</td>
<td>49</td>
<td>1</td>
<td>32</td>
<td>97.0</td>
<td>70.8</td>
<td>39.5</td>
<td>99.2</td>
</tr>
<tr>
<td>BITSEA Problem Index Cutscores versus Elevated CBCL externalizing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Full sample</td>
<td>136</td>
<td>26</td>
<td>6</td>
<td>37</td>
<td>86.0</td>
<td>84.0</td>
<td>58.7</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td>Nonminority</td>
<td>118</td>
<td>44</td>
<td>3</td>
<td>40</td>
<td>93.0</td>
<td>72.8</td>
<td>47.6</td>
<td>97.5</td>
</tr>
<tr>
<td>BITSEA Internalizing Cutscore versus Elevated CBCL internalizing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Full sample</td>
<td>144</td>
<td>24</td>
<td>6</td>
<td>27</td>
<td>81.8</td>
<td>85.7</td>
<td>52.9</td>
<td>96.0</td>
</tr>
<tr>
<td>BITSEA Externalizing Cutscore versus Elevated CBCL externalizing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical</td>
<td>Full sample</td>
<td>149</td>
<td>13</td>
<td>9</td>
<td>34</td>
<td>79.1</td>
<td>92.0</td>
<td>72.3</td>
<td>94.3</td>
</tr>
</tbody>
</table>

Note: BITSEA = Brief Infant–Toddler Social–Emotional Assessment; CBCL = Child Behavior Checklist; NPV = Negative Predictive Value; PPV = Positive Predictive Value.
problems are less prevalent and higher in populations in which problems are more prevalent, such as home-visiting programs for high-risk families. A second implication is that problems may be more severe in this sample than in a nonreferred normative sample. Sensitivity may be lower with less severe problems. Thus, the more generous Of Concern score may be preferred in more normative samples. In a normative epidemiologic sample, the Of Concern score demonstrated acceptable sensitivity (93%), specificity (78%), and PPV (81%) relative to the CBCL (Briggs-Gowan et al., 2004). The consistency of these findings across different samples supports the use of the BITSEA in more normative populations than studied here.

Limitations
Design limitations of this study deserve discussion. First, too few children had psychiatric disorders to support receiver operator curve analyses to test whether alternate cutpoints would optimize sensitivity and specificity. Second, as is often the case with clinical and research work with young children, parent-report measures were used as validators, raising the problem of shared method variance contributing to validity. This is particularly true for comparisons with the CBCL because both measures are checklists. However, the PAPA interview extends beyond simple parent report because the scoring algorithms integrate interviewer ratings of severity, onset, duration, and frequency with empirical data about the normative frequency of the behaviors in young children. Nonetheless, independent measures of functioning and observed behavior would have strengthened the findings. Finally, real-world applications of the BITSEA were not examined. Future research examining how incorporating brief tools (alone or more ideally in combination with clinical follow-up) in screening, surveillance, and home-visiting efforts influences feasibility and affects detection, referral, and treatment outcomes is needed.

Conclusion
Brief measures such as the BITSEA have the potential to support primary and secondary intervention efforts, aiding in the screening, identification, and assessment of an array of clinically significant conditions across the internalizing and externalizing domains. Evidence of validity in assessing internalizing problems is especially promising given that they are often undetected and untreated in young children. Evidence of validity in an ethnically and socioeconomically heterogeneous sample further supports the use of the BITSEA with children from diverse backgrounds. Thus, the BITSEA appears to be an appropriate tool for supporting the efforts of professionals engaged in early identification and assessment in a variety of settings involving young children, including pediatric primary care, screening and surveillance, and the growing number of home-visiting programs for at-risk families.

Funding
This work was supported by the National Institute of Mental Health (grant number R01MH66645). We also thank study participants for making this study possible. The first and second authors receive royalties related to the licensing of the BITSEA.

Conflicts of interest: None declared.


