Screening for Adolescents’ Internalizing Symptoms in Primary Care: Item Response Theory Analysis of the Behavior Health Screen Depression, Anxiety, and Suicidal Risk Scales

Katherine B. Bevans, PhD,* Guy Diamond, PhD,† Suzanne Levy, PhD†

ABSTRACT: Objective: To apply a modern psychometric approach to validate the Behavioral Health Screen (BHS) Depression, Anxiety, and Suicidal Risk Scales among adolescents in primary care. Methods: Psychometric analyses were conducted using data collected from 426 adolescents aged 12 to 21 years (mean = 15.8, SD = 2.2). Rasch-Masters partial credit models were fit to the data to determine whether items supported the comprehensive measurement of internalizing symptoms with minimal gaps and redundancies. Results: Scales were reduced to ensure that they measured singular dimensions of generalized anxiety, depressed affect, and suicidal risk both comprehensively and efficiently. Although gender bias was observed for some depression and anxiety items, differential item functioning did not impact overall subscale scores. Future revisions to the BHS should include additional items that assess low-level internalizing symptoms. Conclusions: The BHS is an accurate and efficient tool for identifying adolescents with internalizing symptoms in primary care settings. Access to psychometrically sound and cost-effective behavioral health screening tools is essential for meeting the increasing demands for adolescent behavioral health screening in primary/ambulatory care.

In recent years, the call for behavioral health (BH) screening in primary and ambulatory care, especially for depression and suicide, has been renewed.1–5 Despite recommendations, barriers to implementation (e.g., lack of physician training, limited time/instrument administration capacity, and difficulty accessing BH services) have limited progress in the adoption of screening for internalizing symptoms in pediatric and adolescent primary care.1–5

A significant barrier to the implementation of adolescent BH screening in primary care is the unavailability of developmentally appropriate and psychometrically sound screening tools.6 Currently, only 2 measures are recommended by the American Academy of Child and Adolescent Psychiatry and the American Academy of Pediatrics1–6: the Beck Depression Inventory-Fast Screen7 and the Patient Health Questionnaire for Adolescents.8 Both tools support the reliable, valid, sensitive, and specific assessment of BH symptoms among adolescents. However, like many other screening tools, these instruments have limitations that hinder their use in primary care settings.9,10

Specifically, best practice guidelines recommend that primary care physicians (PCPs) inquire about the many BH challenges that increase during adolescence, yet most screening tools only assess a single BH dimension (e.g., depression and risk behaviors).7,8 Although screening is recommended to identify depressed youth, failure to assess depressive symptoms within the context of other symptoms and risk factors may be misguided among adolescents for whom comorbidity is the rule, not the exception. In addition, most screening tools fail to use modern information technology to support implementation, scoring, and integration with electronic medical records.11

To address limitations of existing BH screening tools, we developed the Behavioral Health Screen (BHS), a web-based, brief yet comprehensive screening tool for use with adolescents and young adults in primary care settings.12 The BHS assesses both psychiatric symptoms and risk behaviors. BHS items were developed to reflect best practice guidelines in developmental and behavioral pediatrics. A team of 20 national experts reviewed and revised the BHS items, and focus groups were conducted.
with pediatricians to ensure that the instrument is meaningful and feasible for use in primary care settings. A preliminary evaluation using classical test theory (CTT) methods revealed that the BHS Depression, Anxiety, and Suicidal Risk Scales are internally consistent (Cronbach’s alpha = 0.75-0.87) and demonstrate good convergent and divergent validity with standardized measures assessing similar domains ($r^2 = .48-.72$, $p < .0001$). Sensitivity and specificity for the Internalizing Scales are strong with an overall accuracy ranging between 78% and 85%. Using cut scores established through receiver operating curve analyses, patients with increased internalizing symptoms were found to be at least 4 times more likely to indicate that they experienced stressors (e.g., minority sexual preference and violent victimization) and engaged in risk behaviors (e.g., drug and alcohol use and school failure). Given preliminary evidence of its strong psychometric properties, the BHS was adopted in 2 recent studies. Administration of the BHS in an emergency department (ED) was perceived as acceptable to both patients and providers. Of 4018 eligible patients, 1327 (33%) were approached by clinical staff to complete the BHS-ED. A total of 857 (64.6%) patients completed the BHS-ED and 470 (35.4%) refused. Compared with before implementation, there were higher rates of identification of psychiatric problems (10.5% vs 2.5%, odds ratio = 4.58, 95% confidence interval: 3.53-5.94) and more frequent assessments conducted by social workers or psychiatrists (8.3% vs 1.7%). Using the BHS in 10 primary care settings, over 1200 youth were screened in 18 months. Of these, 110 endorsed past or current suicidal ideation. These studies highlight the BHS' potential as a feasible, acceptable, and useful tool for the identification of youth at risk for clinically significant BH problems. Thus, further investigation of the instrument’s psychometric properties is warranted, particularly an in-depth review of its performance when evaluated using modern measurement approaches. CTT methods, those that were previously used to evaluate the BHS, are based on the assumption that respondents’ observed scores are an estimate of their true scores plus or minus some unobservable measurement error. CTT methods are useful in that they depend on few a priori assumptions and produce results that are relatively easy to interpret. However, these methods are limited in that true scores are not exclusively based on respondent characteristics but may also depend on the content of the assessment tool. Therefore, it cannot be determined whether variability in scores is attributable to respondents’ outcome state or properties of the items administered. As a result, it is difficult to compare the outcomes of respondents based on results of different assessment instruments. Moreover, CTT methods assume (often erroneously) that the reliability of a measure remains constant for all respondents regardless of their level on the construct being assessed. In contrast, a unique aspect of modern measurement methods grounded in item response theory (IRT) is that these procedures support the differentiation of respondent and item characteristics through the property of parameter invariance. A cornerstone of all IRT models is the joint measurement of people and items, specifically that a respondent’s “ability” (e.g., level of the underlying BH dimension) and item “difficulty” (how much of the BH problem a person must have to endorse the item) are placed on the same continuum and assessed using the same metric. The joint measurement of respondent and item parameters ensures that IRT procedures are uniquely positioned to provide information about how well a measure assesses the full range of an underlying dimension (e.g., very low levels to very high levels of internalizing symptoms). IRT techniques can be used to enhance the usefulness and efficiency of an assessment tool by identifying gaps in the measurement of the underlying construct (e.g., when there are no or too few items that provide information about respondents along the full range of the BH continuum) and item redundancy (e.g., when too many items provide information about respondents at a specific point on the BH continuum). In addition, IRT provides a natural framework for studying item bias through the comparison of varying subgroup members with comparable levels of BH on the probability of endorsing specific items. Given the benefits of IRT, this article expands on the previously reported psychometric evaluation of the BHS Internalizing Scales by applying a modern measurement approach.

METHODS

Participants

Participants were recruited from Philadelphia-area primary care offices over 18 months. During the recruitment phase, 1836 eligible adolescents visited the practices, of which 1038 were approached about participation and 839 expressed interest and gave contact information. Of these, 770 were called to schedule assessments, but 54 (7%) could not be reached, 100 (13%) were no longer interested, and 190 did not attend their scheduled appointment (24.8%). In the final sample, 426 (55%) completed the assessment, and of these adolescents, 415 (97.4%) had usable data. Remaining cases were removed because they skipped the majority of items and/or had previously completed the assessment. Participating adolescents were aged 12 to 21 years (mean = 15.8, SD = 2.2). The sample was 66.5% female and 77.5% African-American, 10.7% white, 9.7% Hispanic, and 2.1% of another race.

Measures

Behavioral Health Screen Symptom Scales

The Behavioral Health Screen (BHS) is composed of 13 modules: demographics, medical, school, family, safety, substance use, sexuality, nutrition and eating, anxiety, depression, suicidal risk, psychosis, and trauma and abuse. There are 55 core items and 41 additional
items that probe deeper when relevant core items are endorsed. The program displays 1 item at a time and skips out of follow-up items when core items are not endorsed. When completed, the data are scored and a report is immediately printed at a location of the provider’s choice (e.g., nursing station). A summary page displays scaled scores on depression, anxiety, and suicidal risk, as well as urgent risk items (e.g., suicidal ideation/ attempts and sexual abuse), nonurgent risk items (e.g., substance use and self-induced vomiting), and patient strengths (e.g., supportive parent and school success). For this study, we selected the Anxiety, Depression, and Suicidal Risk Scales for further analyses because screening for these BH challenges has been prioritized by the American Academy of Child and Adolescent Psychiatry and the American Academy of Pediatrics.\(^1\) For a more complete description of the BHS and its web platform, see Diamond et al.\(^12\)

Originally, the BHS included 9 items that assessed adolescents’ depressive symptoms within the past 2 weeks (felt down/unhappy/sad/depressed; lost interest in things; change in appetite; change in sleep; felt grouchy/irritable/angry; little energy; trouble making decisions; felt lonely; and felt worthless). Eight items assessed adolescents’ anxiety symptoms within the past 2 weeks (felt restless/keyed-up/on edge; felt worried; doing things repetitively to relax; unpleasant/upsetting thoughts or images; heart pounding or trouble breathing; felt afraid for no obvious reason; felt afraid in social situations; and worried about peer rejection). The Suicidal Risk Scale was composed of 5 items that assessed symptoms across the life span (felt life was not worth living; thought about killing self; made plan to kill self; tried to kill self; and intentionally hurt self). All BHS internalizing symptom items have 3-level response categories for Depression and Anxiety (never, sometimes, and often) and Suicidal Risk (never, in past year, and in past week).

**Procedure**

Written informed parental consent and adolescent assent were obtained for participants aged 12 and 15 years. In accordance with Pennsylvania law and institutional review board approval, adolescents 14 years and older could consent for themselves. The BHS was computer administered and took approximately 8 to 15 minutes to complete. Participants were paid $20 for their participation and those in need of BH services were appropriately referred. All study procedures were approved by the hospital’s institutional review board and conducted by research staff not associated with the development of the BHS.

**Statistical and Psychometric Analyses**

The general characteristics of each item were assessed using traditional descriptive statistics (e.g., response frequency, mean, and SD). Thereafter, we evaluated sets of items that were developed to assess each of the 3 internalizing symptom constructs: Depression, Anxiety, and Suicidal Risk. Item combinations were evaluated to determine whether they met the item response theory (IRT) assumptions of unidimensionality (only 1 dimension is measured by the items that make up a scale) and local independence (when the level of symptom influencing item responses is held constant, responses to any pair of items are statistically independent). We evaluated the unidimensionality of scales by examining item-to-total correlations, estimating internal consistency reliability, and conducting 1-factor confirmatory factor analyses (CFA). If the dimensionality of a scale was in question, exploratory and CFA were combined to identify a meaningful unidimensional construct. Local independence was evaluated by examining residual correlations among items in the 1-factor models.

Once IRT assumptions were confirmed, Rasch-Masters partial credit models were fit to the data and model and item fit determined using Winsteps.\(^16\) For each item, 3 fit statistics were evaluated (infit, outfit, and post hoc estimated item discrimination parameters \([\alpha]\)) to assess the degree to which item responses were predictable and adequately discriminated among adolescents with varying levels of internalizing symptoms. Thereafter, item difficulty \([b]\) and category threshold estimates were inspected to determine whether items supported the comprehensive measurement of BH symptoms with minimal gaps and redundancies. Item-person maps were generated for each scale to illustrate the distribution of BH symptoms and estimates of item difficulty parameters (Fig. 1).

Finally, tests of uniform differential item functioning (DIF) were used to identify systematic errors due to group bias based on gender and age groups (12–14, 15–17, and 18−21 years). Significant DIF contrast values as evidenced by the Mantel-Haenszel significance test indicate that groups of respondents differ on the frequency with which they endorse an item after adjusting for their overall level of BH symptoms.\(^16\) Because it is defined conditional on the underlying variable measured, DIF is not the same as an unconditional difference between groups. Instead, the research question posed in DIF analyses is whether, after controlling for the underlying construct, the response to an item is related to group membership.

**RESULTS**

**Item Descriptive Statistics**

Item missing data rates were 2% or smaller for all items. As is typical for measures of behavioral health (BH) symptoms administered in the general population, many items were positively skewed. However, all response items were endorsed for every question. The largest floor effect was observed for the item, “tried to kill self” (94.2% endorsed “never”). The largest ceiling effects were observed for “had trouble sleeping or slept all the time” (18.4% endorsed “often”).
The 8 anxiety items were subjected to a 1-factor confirmatory factor analysis (CFA) to assess the degree to which the scale met the item response theory (IRT) assumption of unidimensionality. The 1-factor model was a poor fit for the data, which suggests that the items failed to measure a single latent trait (Comparative Fit Index [CFI] = 0.871, standardized root mean square residual [SRMR] = 0.09). Examination of residual correlations revealed that 2 items, both pertaining to social anxiety remained moderately related, even after accounting for variance associated with the overall internalizing trait (r = 0.14). These items were subsequently removed from the scale as they were thought to better assess a secondary dimension of anxiety related to social situations. In addition, 2 items were found to have relatively low factor loadings; they made minimal contributions to the measurement of anxiety (“How often have there been things that you must do over and over again, or have them exactly right in order to relax?” and “Did your heart pound or did you have trouble breathing when you were not exercising?”). Scale unidimensionality and item local independence were supported for the remaining 5 items, and estimated factor loadings were moderate to large and statistically significant (Table 1). These items were found to be adequately internally consistent (α = 0.81).

Similar analyses were conducted to determine whether the 9 depression items met IRT assumptions. The 1-factor model was a poor fit for the data (CFI = 0.824, SRMR = 0.12). Examination of factor loadings suggested that physically oriented depressive symptoms (e.g., change in eating, change in sleeping patterns, and having little energy) and “trouble making decisions” had relatively low factor loadings. After removal of these items, the remaining 5 depression items were found to measure a single dimension. All items had moderate to large factor loadings (Table 2) and were locally independent. The Depression Scale had adequate internal consistency (α = 0.81).

Finally, a 1-factor CFA model containing all 5 suicidal ideation items was a poor fit for the data (CFI = 0.699, SRMR = 0.19), which was primarily attributable to a relatively low factor loading (0.45) for the item, “have you done anything to intentionally hurt yourself?” After removal of this item, the remaining 4 suicidal ideation items were found to measure a single dimension. All items had moderate to large factor loadings (Table 3) and were locally independent. The Suicide Scale had adequate internal consistency (α = 0.78).

### Estimated Rasch Parameters and Model Fit

Once the unidimensionality of scales and local independence of items within scales were established, Rasch-Masters partial credit models were fit separately to the Anxiety, Depression, and Suicidal Risk Data. Item infit/outfit statistics, discrimination parameters (a), difficulty (b), and response category threshold estimates are presented in Tables 1 to 3. All items had satisfactory fit statistics (outfit and infit statistics between 0.7 and 1.4).
and adequately discriminated among children with varying levels of the BH symptoms ($a \geq 0.75$). Graphical depictions of item difficulties and respondents’ levels of BH symptoms on common continua are shown on the item-person maps (Fig. 1). As expected, adolescents’ responses to the BH screen (BHS) items were positively skewed. Items adequately measured the full range of BH symptoms with few gaps in coverage and minimal overlap in response category threshold estimates (Fig. 1).

For the Anxiety Subscale, estimated people ability levels (thetas) ranged from $-4.69$ to $4.98$ (mean $= -2.85$, SD $= 1.45$), with $40.6\%$ of the sample demonstrating floor effects and $0.01\%$ demonstrating ceiling effects. Average item difficulties (deltas) ranged from $-1.20$ to $2.12$ (mean $= 0.00$, SD $= 1.27$) indicating that items covered a broad range in self-reported anxiety. Furthermore, the magnitude of separation between each of the item difficulties exceeded the recommended level of $0.15$ logits indicating that there is minimal redundancy among items. Similar item and scale characteristics were observed for the Depression Subscale. Estimated people ability levels (thetas) ranged from $-4.27$ to $4.27$ (mean $= -2.37$, SD $= 2.03$), with $41.1\%$ of the sample demonstrating floor effects and $1.4\%$ demonstrating ceiling effects. Average item difficulties (deltas) ranged from $-0.70$ to $0.85$ (mean $= 0.00$, SD $= 0.52$) and there was adequate separation between each of the item difficulties ($\geq 0.15$ logits). For the Suicide Subscale, estimated people ability levels (thetas) ranged from $-6.38$ to $3.94$ (mean $= -5.42$, SD $= 1.95$), with $75.9\%$ of the sample demonstrating floor effects and $0\%$ demonstrating ceiling effects. Average item difficulties (deltas) ranged from $-2.25$ to $2.00$ (mean $= 0.00$, SD $= 1.71$), with adequate separation among all items ($<0.15$ logits).

**Differential Item Functioning**

Tests of differential item functioning (DIF) revealed that as expected, there were some systematic differences in the ways that males and females responded to anxiety items. After controlling for the underlying levels of anxiety, females were more likely than males to report feeling worried, whereas males were more likely than females to report feeling restless, keyed-up, or anxious (Table 1). Gender-based DIF was also evidenced for the depression items. After controlling for underlying levels of depression, females were more likely than males to lose interest in things and feel worthless as a person. Conversely, males were more likely than females to report feeling sad and grouchy or irritable (Table 2). Despite evidence for significant gender-based DIF for the aforementioned items, the expected total test scores (based on the mean of all anxiety or depression items as a function of theta) did not differ substantially for males and females. For both subscales, consistencies in expected total test scores by gender are attributable to the
The inclusion of equal numbers of items that are biased against each of the gender groups. The 2 anxiety items with gender DIF are in opposite directions (1 item favors males and the other favors females) and therefore cancel each other out at the test level. A similar effect was observed for depression items with gender DIF; the 2 depression items biased against males are counterbalanced by the 2 items that are biased against females. Gender and age DIF were not observed for the suicidal ideation items.

### Table 2. Descriptive Statistics, Correlations with Total, Factor Loadings, Rasch Model Fit Statistics, Category Thresholds, and DIF Contrasts for Depression Items

<table>
<thead>
<tr>
<th>Domain/Item</th>
<th>Mean (SD)</th>
<th>CFA Factor Loading</th>
<th>Item Infit</th>
<th>Item Outfit</th>
<th>Item (a)</th>
<th>Item (b)</th>
<th>Response Category Thresholds</th>
<th>DIF Contrasts Male vs Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>During the past 2 wk, how often have you been grouchy, irritable, or angry for no reason at all?</td>
<td>1.50 (0.70)</td>
<td>0.65</td>
<td>1.15</td>
<td>1.15</td>
<td>0.77</td>
<td>−0.70</td>
<td>−3.54 −1.13 0.80</td>
<td>.67(^a)</td>
</tr>
<tr>
<td>During the past 2 wk, how often have you felt lonely even when you were around friends or family?</td>
<td>1.44 (0.68)</td>
<td>0.77</td>
<td>0.87</td>
<td>0.88</td>
<td>1.16</td>
<td>−0.34</td>
<td>−3.42 −0.95 1.29</td>
<td>−</td>
</tr>
<tr>
<td>During the past 2 wk, how often have you felt down, unhappy, sad, or depressed most of the day for several days at a time?</td>
<td>1.41 (0.61)</td>
<td>0.80</td>
<td>0.83</td>
<td>0.81</td>
<td>1.22</td>
<td>0.02</td>
<td>−3.50 −0.72 1.63</td>
<td>.66(^a)</td>
</tr>
<tr>
<td>During the past 2 wk, how often have you lost interest in things you used to enjoy?</td>
<td>1.41 (0.60)</td>
<td>0.70</td>
<td>1.07</td>
<td>1.05</td>
<td>0.92</td>
<td>0.17</td>
<td>−3.47 −0.69 1.49</td>
<td>−.56(^a)</td>
</tr>
<tr>
<td>During the past 2 wk, how often have you felt that you were worthless or a failure as a person?</td>
<td>1.25 (0.54)</td>
<td>0.64</td>
<td>1.01</td>
<td>1.19</td>
<td>0.96</td>
<td>0.85</td>
<td>−3.04 −0.36 2.02</td>
<td>−.76(^a)</td>
</tr>
</tbody>
</table>

CFA, confirmatory factor analysis; \(a\), item discrimination parameter; \(b\), item difficulty parameter; DIF, differential item functioning; TLI, Tucker-Lewis Index; RMSEA, Root mean square error of approximation; SRMR, standardized root mean square residual. CFI ≥ 0.97, TLI ≥ 0.95, RMSEA ≤ 0.090 (90% CI = 0.054–0.130), SRMR ≤ 0.02. Response categories: 1 = never, 2 = sometimes, and 3 = often. \(^a\)Significant DIF contrast values \((p < 0.05)\) as evidenced by the Mantel-Haenszel significance test indicate gender differences on frequency of the symptom after adjusting for overall level of depression (negative values indicate males are more likely to endorse item; positive values indicate females are more likely to endorse item).

### Table 3. Descriptive Statistics, Correlations with Total, Factor Loadings, Rasch Model Fit Statistics, Category Thresholds, and DIF Contrasts for Suicidal Risk Items

<table>
<thead>
<tr>
<th>Domain/Item</th>
<th>Mean (SD)</th>
<th>CFA Factor Loading</th>
<th>Item Infit</th>
<th>Item Outfit</th>
<th>Item (a)</th>
<th>Item (b)</th>
<th>Response Category Thresholds</th>
<th>DIF Contrasts Male vs Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Felt that life was not worth living</td>
<td>0.47 (0.97)</td>
<td>0.59</td>
<td>1.34</td>
<td>1.39</td>
<td>0.86</td>
<td>−2.25</td>
<td>−6.25 −2.70 0.04</td>
<td>−</td>
</tr>
<tr>
<td>Thought about killing self</td>
<td>0.34 (0.83)</td>
<td>0.67</td>
<td>0.83</td>
<td>0.85</td>
<td>1.17</td>
<td>−1.00</td>
<td>−6.15 −1.84 1.11</td>
<td>−</td>
</tr>
<tr>
<td>Made a plan to kill self</td>
<td>0.12 (0.53)</td>
<td>0.85</td>
<td>0.71</td>
<td>0.92</td>
<td>1.30</td>
<td>1.24</td>
<td>−5.71 −0.17 3.48</td>
<td>−</td>
</tr>
<tr>
<td>Tried to kill self</td>
<td>0.12 (0.50)</td>
<td>0.79</td>
<td>0.79</td>
<td>1.19</td>
<td>1.12</td>
<td>2.01</td>
<td>−5.73 −0.05 2.47</td>
<td>−</td>
</tr>
</tbody>
</table>

CFA = confirmatory factor analysis; \(a\), item discrimination parameter; \(b\), item difficulty parameter; DIF, differential item functioning; TLI, Tucker-Lewis Index; RMSEA, Root mean square error of approximation; SRMR, standardized root mean square residual. CFI ≥ 0.940, TLI ≥ 0.879, RMSEA ≤ 0.089 (90% CI = 0.062–0.150), SRMR ≤ 0.05. Response categories: 1 = never, 2 = in lifetime, and 3 = in past week.
DISCUSSION

This study provides further support for the Behavioral Health Screen (BHS), a psychometrically robust screening tool for the identification of adolescents with internalizing symptoms. Evaluation of scale dimensionality resulted in retention of items that are core to generalized anxiety and most commonly experienced by adolescents with depressed affect and those at risk for suicidal attempts. Consistent with prior research on anxiety subtypes, symptoms that characterize social anxiety, obsessive-compulsive disorder, and panic disorder were not found to contribute to the single dimension of generalized anxiety as assessed by the BHS.17 Similarly, items that reflected the somatic manifestation of depression failed to contribute to the BHS Depression Scale, which is primarily composed of negative affect items. The self-injury item (e.g., cutting) was removed from the Suicidal Risk Scale because it failed to contribute to the suicidal risk construct, perhaps reflecting the multitude of reasons why adolescents engage in self-injurious behaviors that are unrelated to suicidal ideation.18 Removal of the aforementioned items from the BHS Internalizing Scales is not an attempt to refute their clinical significance and indeed these items are retained on the BHS as “indicator” items that if endorsed are highlighted on the automatically generated report. Rather, they are excluded from the calculation of the BHS Internalizing Symptom Scale scores because they failed to contribute to unidimensional anxiety, depression, and suicidal risk constructs.

We applied item response theory (IRT) to the psychometric evaluation of the BHS Internalizing Scales because of its many advantages over classical test theory (CTT) methods. Most notably, IRT-based procedures support the simultaneous estimation of respondent and item parameters, which supports the identification of ceiling/floor effects and areas along the BH continua for which there are gaps or redundancies in the measurement of symptom severity. Comparison of person and item parameters is best achieved through inspection of the item-person maps (Fig. 1). The maps reveal the expected positive skew in the manifestation of all internalizing symptoms. Given both the importance and relatively low prevalence of adolescents’ BH symptoms, computer-based self-administered screening tools are of substantial value in primary care settings. They support the accurate and expeditious identification of adolescents who require further BH evaluations, while reducing burden on health care providers. Without the support of universal screening tools such as the BHS, health care providers may have difficulty complying with American Academy of Child and Adolescent Psychiatry and American Academy of Pediatrics recommendations for improved BH screening of adolescents (e.g., they may inquire about symptoms using overly general, developmentally inappropriate, and invalidated questions).1

Furthermore, inspection of the item-person maps suggests that the BHS Anxiety Scale adequately measures the full range of generalized symptoms with few gaps in its measurement and minimal item redundancies. Thus, the BHS Anxiety Scale is both comprehensive and efficient. Some gaps were observed in the assessment of Depression and Suicidal Risk Scales. Specifically, future iterations of these scales should include more symptoms endorsed by adolescents with mild depressive symptoms and those at minimal risk for suicidal ideation. In addition, subsequent versions of the BHS may apply computerized adaptive test (CAT) technology, which is made possible by the IRT-based estimation of item parameters. CAT algorithms are applied to “match” the most appropriate set of BHS items to each respondent based on their level of underlying BH symptoms.19 In CAT assessment, the sequence of items administered depends on an individual’s responses to previously administered items. Based on the respondent’s answers to prior questions, only those items that provide the most information about their unique level of BH would be administered. In this way, a smaller number of items could be administered, decreasing administration time without sacrificing measurement precision.

In addition to supporting CAT administration, IRT methods underlie the detection of item bias using the differential item functioning (DIF) framework. The detection of item bias is important for achieving measurement equivalence in heterogeneous populations (e.g., by gender or age groups).10 After controlling for estimates of BH symptoms, boys and girls responded differently to 4 depression items and 2 anxiety items. However, DIF failed to impact the total expected scores on the Depression and Anxiety Scales and therefore may be ignored assuming that scale scores are derived by combining responses to all items. Future iterations of the BHS that apply CAT methods for which a subset of items are administered for each scale will require alternative approaches to ensuring measurement equivalence by accounting for DIF. In 1 such approach, item parameters estimated from the whole sample would be specified for DIF-free items and items with DIF would have parameters estimated separately in different demographic groups.20

The BHS is a valuable tool for clinical research and care applications. BHS Internalizing Scales reliably and accurately measure unidimensional depression, anxiety, and suicidal risk dimensions. Therefore, each scale can be used in isolation to measure primary and/or secondary endpoints in clinical or comparative effectiveness research. Although the BHS Internalizing Scales assess unidimensional BH constructs, the symptoms often congregate in patient clinical presentation. Nearly 60% of suicidal youth are depressed, and youth with comorbid depression and anxiety are often the most difficulty to treat.21 Thus, in clinical contexts, the BHS’s value lies in its capacity to measure multiple BH dimensions with relatively few items.
Study Limitations

To date, the BHS has been evaluated using population-based approaches and additional study is needed to confirm it has adequate predictive validity among individuals assessed in clinical settings. This study was conducted using a convenience sample, which was primarily composed of African-American inner city youth. Replication of these analyses with a larger, more culturally diverse sample will help improve our confidence in the generalizability of findings.

CONCLUSION

Many view primary care as the center of a clinically and cost-effective health care delivery system. For many patients, primary care physicians (PCPs) are the point of entry into medical services for a variety of mental health-related problems. Unfortunately, the PCP is often overwhelmed with too many demands and too little time to accomplish this. This reality raises concerns that adding mental health assessment to their responsibilities will further burden busy providers. However, PCPs already encounter high rates of patients with mental health problems and have little resources to manage these kinds of disorders. Providing the PCP with mental health tools that are clinically relevant and time efficient is imperative. Although we need a broader system change approach to mental health in primary care, access to reliable, valid, and meaningful screening tools such as the Behavioral Health Screen (BHS) must be part of the solution.

ACKNOWLEDGMENTS

We thank all the referring medical staff, social workers, the participating families, and all the students and volunteers who gave invaluable time and effort. We also thank the following members for their administrative, technical, or material support: Jennifer Barnett, Jacqueline Reynier, Allison Reany, Anna Westlin, Jacqueline Spitzer, Erin Lucey, Matthew Haines, Stephen DeLuca, Colleen Walsh, Ruth Craig, Jessica Hamilton, and Alana O’Malley.

REFERENCES


Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.