Journal of Consulting and Clinical Psychology
2010, Vol. 78, No. 4, 526–536

0022-006X/10/$12.00 DOI: 10.1037/a0019602

Evaluation of the Brief Problem Checklist: Child and Caregiver Interviews to Measure Clinical Progress

Bruce F. Chorpita and Steven Reise
University of California, Los Angeles

John R. Weisz
Harvard University and Judge Baker Children's Center, Boston, Massachusetts

Kathleen Grubbs
University of Hawaii, Manoa

Kimberly D. Becker
Johns Hopkins Bloomberg School of Public Health

Jennifer L. Krull
University of California, Los Angeles

The Research Network on Youth Mental Health

Objective: To support ongoing monitoring of child response during treatment, we sought to develop a brief, easily administered, clinically relevant, and psychometrically sound measure. ***Method:*** We first developed child and caregiver forms of a 12-item Brief Problem Checklist (BPC) interview by applying item response theory and factor analysis to Youth Self-Report (YSR; Achenbach & Rescorla, 2001) and Child Behavior Checklist (CBCL;Achenbach & Rescorla, 2001) data for a sample of 2,332 youths. These interviews were then administered weekly via telephone to an ethnically diverse clinical sample of 184 boys and girls 7–13 years of age and their caregivers participating in outpatient treatment, to examine psychometric properties and feasibility. ***Results:*** Internal consistency and test–retest reliability were excellent, and factor analysis yielded 1 internalizing and 1 externalizing factor. Validity tests showed large and significant correlations with corresponding scales on paper-and-pencil administrations of the CBCL and YSR as well as with diagnoses obtained from a structured diagnostic interview. Discriminant validity of the BPC interviews was supported by low correlations with divergent criteria. Longitudinal data for the initial 6 months of treatment demonstrated that the BPC significantly predicted change on related measures of child symptoms. Estimates obtained from random coefficient growth models showed generally higher slope reliabilities for the BPC given weekly relative to the CBCL and YSR given every 3 months. ***Conclusions:*** Given their combination of brevity and psychometric strength, the child and caregiver BPC interviews appear to be a promising strategy for efficient, ongoing assessment of clinical progress during the course of treatment.

Keywords: assessment, brief, item response theory, outcomes, measurement

Movement toward scientifically driven decision making in mental health practice requires robust data collection methods that are supported by empirical evidence (e.g., Achenbach, 2005; Brooks & Kutcher, 2003; Hunsley & Mash, 2007; Klein, Dougherty, & Olino, 2005; Meyer et al., 2001). Hunsley and Mash (2007) stated

> it is important not to overlook the interplay between assessment and intervention that is at the heart of providing evidence-based psychological treatments. This assessment-intervention dialectic, involving the use of assessment data both to plan treatment and to modify the treatment in response to changes in a client's functioning and goals [Weisz, Chu, & Polo, 2004], means that evidence-based assessment has relevance for a broad range of clinical services. (p. 46)

This comment underscores the importance of investigation and continued development of psychometrically sound data collection instruments and procedures that have the potential to be widely utilized, both in research and in practice. Attempts to develop

Bruce F. Chorpita, Steven Reise, and Jennifer L. Krull, Department of Psychology, University of California, Los Angeles; John R. Weisz, Department of Psychology, Harvard University, and Judge Baker Children's Center, Boston, Massachusetts; Kathleen Grubbs, Department of Psychology, University of Hawaii, Manoa; Kimberly D. Becker, Johns Hopkins Bloomberg School of Public Health; The Research Network on Youth Mental Health. The Research Network on Youth Mental Health is a collaborative network funded by the John D. and Catherine T. MacArthur Foundation. Network members at the time of this work (in alphabetical order) included the following: Bruce F. Chorpita, Robert Gibbons, Charles Glisson, Evelyn Polk Green, Kimberly Hoagwood, Kelly Kelleher, John Landsverk, Stephen Mayberg, Jeanne Miranda, Lawrence Palinkas, Sonja Schoenwald, and John R. Weisz (network director).

This work was supported by grants from the John D. and Catherine T. MacArthur Foundation to Bruce F. Chorpita and John R. Weisz, from the Annie E. Casey Foundation to Bruce F. Chorpita, and from the Norlien Foundation and the National Institute of Mental Health (Grants MH068806 and MH085963) to John R. Weisz.

Correspondence concerning this article should be addressed to Bruce F. Chorpita, Department of Psychology, Franz Hall 3227, Box 951563, University of California, Los Angeles, 90095-1563. E-mail: chorpita@ucla.edu

empirically supported measurement methods have thus understandably been a focus of research for several decades and are now central to the current emphasis in our field on evidence-based assessment as a partner to evidence-based intervention.

Increasingly, in both research and routine clinical care, there is a need for measures of treatment progress and outcome that can be used frequently to monitor client response over the course of treatment and to inform ongoing treatment planning and case supervision (cf. Chamberlain & Reid, 1987; Webster-Stratton & Spitzer, 1991). In research, the need for frequent measurement is underscored by three emerging trends: (a) the increasing use of statistical procedures, such as hierarchical linear modeling (e.g., Raudenbush & Bryk, 2002), which can leverage measurements across multiple time points to evaluate trajectories rather than merely the degree of change; (b) observations that significant clinical improvements can occur early in treatment (e.g., Ilardi & Craighead, 1994), such that measures taken only at posttreatment may underestimate or even fail to detect meaningful differences between groups producing different rates of clinical change (see, e.g., Weisz et al., 2009); and (c) the increasing use of randomized controlled trial designs comparing evidence-based treatments with usual care of uncontrolled duration and dose (see Weisz, Jensen-Doss, & Hawley, 2006)—designs in which rate of change provides less biased assessment of outcome than posttreatment measurement alone.

In clinical practice, the need for frequent measurement derives partly from the fact that the clinician's ability to plan and adjust treatment procedures over the course of a treatment episode is enhanced to the extent that there is an ongoing flow of information about how the client is responding (e.g., Chorpita, Bernstein, Daleiden, & the Research Network on Youth Mental Health, 2008; Lambert, Harmon, Slade, Whipple, & Hawkins, 2005). In both research trials and routine clinical care, implementing frequent ongoing measurement can help address another problem as well: Clients may stop treatment at any point without prior announcement and thus may not be available for posttreatment outcome assessment. In cases in which a recurrent assessment strategy is in place throughout treatment, the "final outcome" assessment is implicitly the last of the recurrent assessments, and trajectory-of-change data are automatically available for the full episode of care.

Whereas the need for frequent assessment throughout treatment seems clear in both clinical research and clinical care contexts, the available measures, for the most part, have two limitations: (a) they tend to be narrow in focus—specific to a particular disorder or problem cluster (e.g., depression, anxiety, conduct)—such that they cannot be aggregated or even administered across a large number of youths within an organization or system, and/or (b) they tend to be too time consuming for frequent use. Thus, what is needed are measures that are (a) general enough to be relevant to an array of treated conditions and thus suitable for the diversity of cases and the comorbidity that are so often seen in clinically representative trials and in routine clinical care; (b) brief enough to be used repeatedly without imposing excessive measurement burden; and (c) psychometrically sound and correlated with lengthier, more established measures of psychopathology and/or functioning.

Unfortunately, measurement breadth and quality are goals that can often compete directly with measurement efficiency. That is, all other things being equal, longer measures generally meet these goals better than shorter ones. However, in some contexts, attempts to resolve these conflicting aims through item response theory (IRT) have proven helpful (e.g., Ackerman, Gierl, & Walker, 2003; Embretson & Reise, 2000). IRT as an analytic strategy can be used to design tests whose items provide the maximum information about a respondent's true score by eliminating less informative and redundant items, often producing briefer measures that maintain or approximate the psychometric properties of the longer measures from which they were derived. IRT has been applied with varying degrees of success in areas such as achievement and ability testing (e.g., Ackerman et al., 2003) and health assessment (e.g., Ware, Gandek, Sinclair, & Bjorner, 2005). Despite its promise, the use of IRT to balance breadth, efficiency, and quality has not been very evident in the area of child psychopathology assessment.

In this article, we describe an application of IRT to assessment in child psychotherapy. We report the psychometric properties of two measures, the Brief Problem Checklist (BPC) child and caregiver interviews, that were created using factor analysis and IRT as applied to items from two longer, widely used evidence-based instruments: the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001) and the Youth Self-Report (YSR; Achenbach & Rescorla, 2001), which make up the Achenbach System for Empirically Based Assessment (ASEBA). The BPC includes 12 items designed for brief interview administration. By virtue of its efficiency, the BPC creates the opportunity for repeated periodic assessments and for clinicians to receive feedback about clinical progress over the course of treatment—even in the absence of treatment sessions, such as when families skip several sessions because of illness or discontinue services altogether without a formal termination.

To assess the empirical potential of the BPC measures, we investigated their psychometric properties as well as their efficiency. We predicted that the BPC child and caregiver interviews would yield not only a Total Problems score but also two factorially valid scales, with items corresponding to "Internalizing" and "Externalizing" constructs. The scales were also predicted to show acceptable internal consistency and test–retest reliability, and it was expected that BPC child and caregiver interview scales would show good convergent and discriminant validity in relation to relevant scales of the CBCL and YSR. We also expected the BPC interview scales to predict changes in full CBCL and YSR scales at 3 and 6 months. Finally, we report on the time and effort required to administer these interviews as an index of their measurement efficiency.

Method

Participants

Participants were 184 children who were assessed and offered treatment for problems with anxiety, depression, or disruptive behavior. Nine different school-based and community agencies across two cities participated in the investigation, and treatment was delivered by 78 therapists with the following training: 64% master's level, 35% doctoral level, and 1% bachelor's level. Therapists reported the following orientations: cognitive behavioral (38%), eclectic (23%), psychodynamic (15%), behavioral (8%), family systems (8%), and other (8%; e.g., Adlerian, health realization). Treatment involved child individual, parent individual, or

family therapy formats and took an average of 221.9 days (*Mdn* = 198; *SD* = 143.7). The mean number of sessions was 16.8 (*Mdn* = 14; *SD* = 11.4).

Participant inclusion criteria were as follows: (a) primary diagnoses of an anxiety disorder (or selective mutism), depressive disorder, or disruptive behavior disorder, or parent- or child-reported disturbances of anxiety, depression, or disruptive behavior that did not meet criteria for clinical diagnosis; (b) significant elevation (T > 64) on at least one of the Internalizing or Externalizing narrowband scales of the CBCL (description below) or YSR (description below) (e.g., CBCL Anxious/Depressed scale T score); (c) 7–13 years of age at the time of initial assessment; and (d) primary language of English for parent and child. Of the children meeting inclusion criteria, those with recent psychiatric hospitalizations or suicide attempts or with evidence of psychosis or pervasive developmental disorders at the time of assessment were excluded. The assessments were performed in mental health clinics and school-based behavioral health settings in cities and suburban areas of two U.S. states by trained members of the investigator team.

Grade level ranged from 1 to 9, mean age was 10.64 years (*SD* = 1.80; range = 7.15–13.97), and 127 of the 184 participants were boys (69.0%). Ethnicities reported were Caucasian (*n* = 81; 44.0%), multiethnic (*n* = 58; 31.5%), African American (*n* = 19; 10.3%), Hispanic or Latino/a (*n* = 12; 6.5%), Asian American (*n* = 7; 3.8%), Pacific Islander (*n* = 4; 2.2%), and other (*n* = 3; 1.6%). Information about the primary diagnoses and treatment targets of participants appears in Table 1, and assessment procedures are described below. Diagnostic comorbidity—that is, more than one *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) diagnosis—was evident in 80.8% of the sample (mean number of diagnoses = 2.58; *SD* = 1.40). Most common types of diagnoses (primary or comorbid) were as follows: disruptive behavior disorders (e.g., conduct disorder, oppositional defiant disorder), present in 63.6% of cases; attention-deficit/hyperactivity disorder, present in 56.8% of cases; one or more anxiety disorders, present in 53.4% of cases; and mood disorders (e.g., major depressive disorder, dysthymic disorder), present in 46.6% of cases.

Table 1
Primary Diagnoses or Subclinical Primary Problems of the Sample

Principal diagnosis or problem	Frequency	
	n	%
Separation anxiety disorder	23	12.50
Generalized anxiety disorder	19	10.33
Social anxiety disorder	8	4.35
Obsessive–compulsive disorder	4	2.17
Selective mutism	2	1.09
No diagnosis: Subclinical anxiety	6	3.26
Anxiety-related problems Total	62	33.70
Major depressive disorder	22	11.96
Depressive disorder NOS	6	3.26
Dysthymic disorder	4	2.17
Adjustment disorder with depressed mood	1	0.54
No diagnosis: Subclinical depressed mood	4	2.17
Depression-related problems Total	37	20.11
Oppositional defiant disorder	53	28.80
Conduct disorder	24	13.04
Disruptive behavior disorder NOS	1	0.54
Adjustment disorder with disturbance of conduct	1	0.54
No diagnosis: Subclinical disruptive behavior	6	3.26
Disruptive-behavior-related problems Total	85	46.20
Total	184	100.00

Note. NOS = not otherwise specified.

Informants for the caregiver interviews were mothers (biological, adoptive, or step-mothers; *n* = 153; 85.5%), grandparents (*n* = 15; 8.4%), fathers (*n* = 9; 5.0%), uncle (*n* = 1; 0.6%), and great-great aunt (*n* = 1; 0.6%). Parental marital status in participant families was as follows: married, 39.1%; divorced, 22.3%; single parent, 16.3%; separated, 7.6%; living with partner, 6.0%; and widowed, 5.4%. Modal education level for parents was a high school diploma or equivalent. Household income was assessed with a checklist of ranges that spanned $20,000 each. The median income was in the $20,000–$39,000 range (56.0% of the sample fell within or below this category). Household income supported 3.80 family members on average (*SD* = 1.43).

Measures

Children's Interview for Psychiatric Syndromes (ChIPS) and Children's Interview for Psychiatric Syndromes–Parent Version (P-ChIPS). The ChIPS (Weller, Weller, Rooney, & Fristad, 1999a) and P-ChIPS (Weller, Weller, Rooney, & Fristad, 1999b) are structured interviews designed to be administered to children 6–18 years of age and their parents, respectively, by trained interviewers. Both ChIPS and P-ChIPS assess 20 different *DSM–IV* Axis I disorders as well as psychosocial stressors. Questions use simple language and short sentence structure to enhance participant comprehension and cooperation. The ChIPS has shown moderate agreement with the Schedule for Affective Disorders and Schizophrenia for School-Age Children (Swenson et al., 2007) and high levels of agreement with the Diagnostic Interview for Children and Adolescents—Revised–Child Version (Reich, Shayka, & Taibleson, 1991; Rooney, Fristad, Weller, & Weller, 1999) using kappa coefficients for agreement between interviews. Good interrater reliability and test–retest of both the ChIPS and P-ChIPS have also been demonstrated in previous studies in both clinical and community samples (Fristad, Cummins, et al., 1998; Fristad, Glickman, et al., 1998; Fristad, Teare, Weller, Weller, & Salmon, 1998; Teare, Fristad, Weller, Weller, & Salmon, 1998a, 1998b).

CBCL. The CBCL (Achenbach & Rescorla, 2001) is a questionnaire on which parents rate the extent to which their child shows each of 118 behavioral and emotional problems. Each item is rated on a 0–2 scale, with higher scores reflecting higher problem levels. The problem items on the CBCL can be summed to yield eight narrow-band syndrome scales (Anxiety/Depression, Withdrawn/Depressed, Somatic Complaints, Attention Problems, Thought Problems, Social Problems, Aggression, and Delinquent Behavior), two broad-band syndrome scales (Internalizing and Externalizing), and a Total Problems scale. Validity and reliability of the syndrome scales are excellent (Achenbach & Rescorla, 2001), and extensive normative data are available for children ranging in age from 6 to 18 years. More recently, Achenbach, Dumenci, and Rescorla (2003) developed *DSM*-oriented scales

from a subset of CBCL items, on the basis of expert assignment of selected items to dimensions of *DSM–IV* diagnostic groupings. These scales include the following: Anxiety Problems, Affective Problems, Somatic Problems, Attention-Deficit/Hyperactivity Problems, Oppositional Defiant Problems, and Conduct Problems.

YSR. The YSR (Achenbach & Rescorla, 2001) is a youth-report measure corresponding to the CBCL. The problem section of the YSR has 118 items that form the same 10 syndrome scales and total score generated by the CBCL. Independent normative data are available for the YSR, and the measure has similarly strong psychometric properties. In addition, the YSR also has newly developed *DSM* scales, corresponding to those on the CBCL (Achenbach et al., 2003).

BPC: Child and caregiver interviews. The BPC interviews are two 12-item measures, one for child informant and one for caregiver informant, adapted from items on the CBCL and YSR (Achenbach & Rescorla, 2001) for an orally administered interview designed to track clinical outcomes over time. These measures were designed to yield two scales: Internalizing and Externalizing, which were intended to correspond to the targets of treatment for many children and adolescents. The scale scores of the BPC interviews are based on the raw sum of item responses, each of which ranges from 0 to 2. Thus, scores on the six-item Internalizing and Externalizing scales each range from 0 to 12, and scores on the Total Problems scale range from 0 to 24, with higher scores indicating increased problem levels.

Test construction involved the application of factor analysis and IRT to CBCL and YSR item-level data from an archival clinical sample of boys and girls 8–12 years of age (n = 2,332). As expected, common factor analyses of parent- and child-reported items from the Anxious/Depressed, Withdrawn/Depressed, Rule Breaking, and Aggressive scales using an oblique (promax) rotation each produced two factors corresponding to internalizing and externalizing dimensions. Because we intended to use consistent item sets across informants for the final measure, we selected items that had high factor loadings on both the CBCL and YSR, which yielded 14 internalizing items (anxious, cries, fears impulses, feels guilty, feels persecuted, feels unloved, feels worthless, has to be perfect, lonely, nervous, sad, self-conscious, suspicious, worrying) and 20 externalizing items (argues, attacks others, brags, cruelty to others, demands attention, destroys others' property, destroys own property, disobedient at home, disobedient at school, fights a lot, jealous, loud, moody, screams, shows off, stubborn, talks too much, teases, temper tantrums, threatens people).

We then used MULTILOG Version 7.0.3 to fit Samejima's (1969) graded response model to the data, using marginal maximum likelihood estimation, which produced category response curves (how the probability of a category response changes as a function of the trait), item information curves (how discriminating an item is at different levels of the trait), and estimated IRT parameters. We used these results to select items that maximized information functions corresponding to the latent trait values ranging from the nonclinical mean to two standard deviations above the mean. Thus, items were selected specifically to be maximally and nonredundantly sensitive to clinical change in the range from approximately the 50th to the 95th percentile on a nonclinical distribution. For example, items such as "attacks others" (externalizing) and "suspicious" (internalizing) were discarded because their item location parameters suggested that they discriminated respondents only at the very high end of their respective latent traits. In the same manner, items such as "demands attention" (externalizing) and "nervous" (internalizing) were discarded because of their positions at the very low end of their respective latent traits. From the remaining items, we then selected selecting those items with the highest discrimination parameters across child and caregiver results (e.g., the item "cries," which apparently correlated with temper tantrums as well as depressed mood, was not selected because of its very low discrimination parameter for the Internalizing scale). Items were then evaluated in sets of six to see which produced the best performing information functions. When two items had quite similar IRT parameters (e.g., "cruel" and "destroys things"), only one item was chosen at a time for inclusion in a candidate set. Because seven items performed well on the Externalizing scale, one item on the child interview and one item on the parent interview were ultimately combined from two items from the YSR and CBCL, respectively (e.g., "disobedient at home" and "disobedient at school" were changed to "disobedient at home or school"). Once the final sets of six items for each scale were identified, the measure's instructions were reworded to account for oral administration, and the time frame was changed from 6 months to "in the last week."

Procedure

At the time of the initial assessment, child and parent reviewed and signed consent forms describing the study procedures. Child and parent, seen separately, completed the ChIPS and P-ChIPS interviews, respectively, and the study questionnaires, assisted by project staff. Following the initial assessment, the assessor and a doctoral-level assessment supervisor assigned diagnoses according to ChIPS standard procedures. All BPC interviews were administered by phone.

Children entering treatment were then assigned a single phone caller (from a pool of 32), who contacted and interviewed the child and caregiver by phone with a target interval of every 7 days. Interviews were conducted with children and parents separately, and sequencing of the administration was not experimentally controlled but rather was dictated by the availability of each respondent and their preferences for call times (e.g., could be on separate days). Phone callers were blind to all clinical and study-related information about the child participants (e.g., diagnoses) and were only aware of the name, age, and gender of the children. BPC interviews were administered for the full duration of each child's treatment.

Analytic Plan

In the sample of 184 children entering clinical treatment, we first conducted exploratory factor analyses to determine whether the items selected to define the Internalizing and Externalizing scales of the BPC organized themselves according to the expected factor structure. Reliability of the scales was then estimated using two different strategies. From the same initial BPC administration, internal consistency was examined for the BPC Internalizing, Externalizing, and Total Problems scales using Cronbach's alpha coefficient (Cronbach, 1951), after which Pearson product-moment correlations were calculated to assess test–retest stability. Whenever possible, we repeated

analyses from data taken at baseline using data from 3 months post baseline, to estimate whether the psychometric properties of the BPC scales changed over time.

To evaluate validity of the scales, we calculated Pearson correlations between the BPC child and caregiver scales, on the one hand, and selected CBCL and YSR broadband, narrowband, and *DSM* scales, on the other hand, using raw scores from all scales, with the BPC items excluded from the CBCL and YSR scales (whose reliability [internal consistency] coefficients remained high after these adjustments; range = .81–.94). For comparison purposes on convergent validity tests only, we also computed the correlations between BPC scales and CBCL and YSR scales with the overlapping items retained. Alpha level for tests of whether correlations were significantly different from zero was set at .01. To provide an additional test of validity using more independent criteria, we also examined the convergence of BPC scales with *DSM* diagnosis obtained from structured interviews. In addition, we report cross-informant correlations for the BPC scales to demonstrate agreement between caregiver and child reports.

Given that the discriminant and convergent validity criteria in this study were not orthogonal (e.g., YSR Internalizing and Externalizing scales have been significantly correlated in previous studies, and YSR Internalizing and Externalizing scales correlated .58 in this sample), we did not expect discriminant validity coefficients to be zero. To address this issue, we followed our initial zero-order calculation of discriminant validity correlations with separate tests controlling for nontarget variance in the discriminant criteria. For example, when the BPC Child Externalizing score was subjected to discriminant validity tests, the discriminant validity criterion (e.g., YSR Internalizing Total score) was first regressed on YSR Externalizing Total score, and the residuals of this regression (in this example, representing "externalizing-free" internalizing criterion scores) were used as the "adjusted" discriminant validity criterion.

To examine discriminant validity in another way, we also determined whether zero-order correlations of the BPC Internalizing and Externalizing scales were significantly higher for the convergent criteria relative to the discriminant criteria. These tests were conducted using z tests as recommended by Meng, Rosenthal, and Rubin (1992) for identifying whether dependent correlations (i.e., those taken on different measures in the same sample) are significantly different in magnitude. If correlations of the BPC scales with convergent criteria were significantly higher than correlations with discriminant criteria, this would support the validity of the BPC scales.

To examine the ability of the BPC scales to assess change over time, we compared the within-subject regression lines (predicting BPC scores using time as a log function of days) created by the BPC data with changes from baseline in CBCL and YSR scores at 3 and 6 months. We also compared the BPC longitudinal data with the CBCL and YSR longitudinal data (baseline and 3 months) in terms of their ability to predict CBCL and YSR scores at 6 months. Finally, we gathered information on a subset of calls to determine both the level of effort required to gather information (i.e., number of calls required to complete an assessment) and the amount of time required to administer each successful call.

Results

Factor Analysis

Child interview. The 12 items from the child interview were first subjected to exploratory factor analysis using maximum likelihood estimation. The scree plot suggested a two-factor solution (these factors explaining 27.8% and 14.6% of the variance, respectively), consistent with the test design, so two factors were first extracted and then rotated. An oblique (promax) rotation was used because correlated factors were hypothesized. These two factors corresponded to the Externalizing and Internalizing scales that were part of the interview design (see Table 2 for rotated factor loadings). The two factors were correlated .39 with one another. (The full YSR Internalizing and Externalizing scales were correlated .53 with one another in this sample, and Achenbach & Rescorla, 2001, reported a correlation of .58.)

Caregiver interview. The 12 items from the Caregiver interview were also subjected to exploratory factor analysis using maximum likelihood estimation. The scree plot again suggested a two-factor solution (explaining 34.7% and 19.5% of the variance, respectively); thus, two factors were extracted and subjected to promax rotation. These two factors once again corresponded to the Externalizing and Internalizing scales that were part of the interview design (see Table 2 for rotated factor loadings). The relative magnitude of loadings was higher overall than for the child interview, and the ranking of loadings was also slightly different. The factors on the caregiver interview were intercorrelated at .31. (The full CBCL Internalizing and Externalizing scales were correlated .37 with one another in this sample, and Achenbach & Rescorla, 2001, reported a correlation of .54.)

Internal Consistency

All six of the BPC scales were found to have good internal consistency in this sample. Alpha coefficients for the initial child interview ($N = 184$) were as follows: $\alpha_{Internalizing} = .72$; $\alpha_{Externalizing} = .70$; $\alpha_{Total} = .76$. In addition, alpha coefficients for the initial

Table 2
Item Loadings From the Exploratory Factor Analysis of the Child and Caregiver Interviews

	Child		Caregiver	
Item	Internalizing	Externalizing	Internalizing	Externalizing
Worries	.78		.83	
Sad	.63		.56	
Self-conscious	.59		.54	
Worthless	.45		.63	
Fearful	.38		.65	
Guilty	.37		.66	
Disobedient		.77		.77
Tantrums		.62		.81
Argues		.54		.75
Stubborn		.45		.76
Threatens		.42		.38
Destroys		.33		.50

Note. Items are listed in descending order of child loadings within factors. Loadings with values less than .30 are not shown.

parent interview ($N = 184$) were as follows: $\alpha_{Internalizing} = .83$; $\alpha_{Externalizing} = .81$; $\alpha_{Total} = .82$. Because alpha is known to be lower for scales with small numbers of items (Clark & Watson, 1995), we also calculated average interitem correlations for each scale. Clark and Watson (1995) stated that values in the .15–.50 range show that items assess a relatively narrow construct. Average interitem correlations for the child interview were as follows: $r_{Internalizing} = .30$; $r_{Externalizing} = .27$; $r_{Total} = .20$. In addition, average interitem correlations for the parent interview were as follows: $r_{Internalizing} = .42$; $r_{Externalizing} = .45$; $r_{Total} = .28$. These estimates were also obtained on the portion of the sample having data at 3 months after initiating treatment ($n = 133$ for child; $n = 133$ for caregiver), and all measures of internal consistency were slightly higher (e.g., child: $\alpha_{Internalizing} = .84$; $\alpha_{Externalizing} = .74$; $\alpha_{Total} = .83$; parent interview: $\alpha_{Internalizing} = .88$; $\alpha_{Externalizing} = .85$; $\alpha_{Total} = .87$).

Test–Retest Reliability

Next, we assessed test–retest reliability, using participants who had data from both the first and second phone interview within informants. The average retest interval for the 181 cases on the BPC child interview was 8.72 days ($SD = 3.50$; range = 4–29), and the average retest interval for the 180 cases on the BPC caregiver interview was 8.38 days ($SD = 3.01$; range = 1–26). The test–retest coefficients, shown in Table 3, ranged from .72 to .79 at the first two administrations and were somewhat higher at the 3-month period.

Correlations With Criterion Measures

To examine the convergent and discriminant validity of the BPC scales, we examined their correlations with corresponding scales from the YSR and CBCL within each informant, again from the first administration of each instrument. BPC interviews began on average 13.86 days ($SD = 6.56$) after the intake measures were administered. Participants were excluded if their intertest interval was greater than 30 days (19 child reports and 18 caregiver reports were excluded, and those cases did not differ from the rest of the sample on demographic variables). It should be noted that that all estimates of convergent and discriminant validity were subject to a separation in time between administrations of the BPC and its criterion measure (i.e., YSR or CBCL). Thus, the best estimate of the upper limit of a convergent validity coefficient was not 1.0 but the lesser of either scale's test–retest reliability.

Table 3
Test–Retest Reliability Estimates for the Brief Problem Checklist Child and Caregiver Scales at Intake and Three Months After Intake

	Child		Caregiver	
Time period	*r*	*n*	*r*	*n*
Intake				
Internalizing	.76	181	.76	180
Externalizing	.72	181	.78	180
Total	.79	181	.76	180
3 months				
Internalizing	.89	133	.82	133
Externalizing	.79	133	.73	133
Total	.89	133	.76	133

BPC child scales. Table 4 shows the correlations of the BPC Child Internalizing and Externalizing scales with criterion measures. As expected, each BPC scale showed a high and significant correlation with its corresponding scale on the YSR, with all three coefficients at .60 or above (see main diagonal of first three rows of Table 4). Likewise, convergent validity correlations of the BPC Child Internalizing and Externalizing scales and the YSR syndrome scales (rows 4–7) were all moderate to high and statistically significant. Finally, convergent validity coefficients of the BPC Internalizing and Externalizing scales with the YSR *DSM* scales (rows 8–11) were uniformly high and statistically significant. Convergent validity coefficients calculated at 3 months post intake ($n = 131$) were somewhat higher on broadband scales (i.e., $r_{Internalizing} = .78$; $r_{Externalizing} = .81$; $r_{Total} = .76$) as well as on syndrome scales (*r*s ranged from .56 to .84) and *DSM* scales (*r*s ranged from .65 to .84).

Discriminant validity coefficients (zero-order correlations) of BPC Internalizing and Externalizing scales with YSR scales were generally small to moderate in size, but most were statistically significant. After adjustment for nontarget variance, all discriminant coefficients were nonsignificant, suggesting that significance among zero-order discriminant validity coefficients was related to criterion association among YSR scales. Within each grouping of criterion scales (e.g., rows 1 and 2, rows 4–7, and rows 8–11), we compared every convergent validity coefficient with every zero-order discriminant validity coefficient separately for the BPC Internalizing and Externalizing scales (e.g., BPC Internalizing convergent correlations with YSR Anxious Depressed and Withdrawn Depressed tested relative to BPC Internalizing discriminant correlations with YSR Rule Breaking and Aggressive). In all cases, convergent validity coefficients were significantly higher than the discriminant coefficients ($p < .05$). Using data from 3 months post intake ($n = 131$), we found that discriminant coefficients were generally higher than at intake (*r*s ranged from .23 to .40) but in all cases were significantly lower than their convergent counterparts at 3 months.

BPC caregiver scales. Correlations with broad-band CBCL scales showed good convergent validity for all three BPC scales (all greater than .50 and statistically significant; see Table 5). Likewise, correlations of the BPC Internalizing and Externalizing with narrow-band CBCL syndrome and *DSM* scales yielded observations consistent with predictions, correlating significantly and moderately or highly with convergent criteria. Convergent validity coefficients from 3 months post intake ($n = 132$) were again somewhat higher than at intake on broadband scales (i.e., $r_{Internalizing} = .73$; $r_{Externalizing} = .73$; $r_{Total} = .70$), syndrome scales (*r*s ranged from .54 to .73), and *DSM* scales (*r*s ranged from .65 to .71).

Discriminant validity coefficients (zero-order correlations) of BPC Internalizing and Externalizing scales with CBCL scales were generally small in size, and only one was statistically significant. After adjustment for nontarget variance, all discriminant coefficients were nonsignificant, suggesting that the significant zero-order discriminant validity coefficient (between BPC Externalizing and CBCL *DSM* Affective) was due to criterion association between the CBCL *DSM* Affective scale and the CBCL Externalizing scale. Once again, within each grouping of criterion scales (e.g., rows 1 and 2, rows 4–7, and rows 8–11), we com-

Table 4
Correlations Among Brief Problem Checklist (BPC) Child Scales and Selected Scales From the Youth Self-Report (N = 166)

Youth self-report	BPC CHILD Internalizing	BPC CHILD Externalizing	BPC CHILD Total
Internalizing	**.56 [.61]**	**.31** (−.02)	**.54**
Externalizing	**.27** (−.14)	**.50 [.61]**	**.45**
Total	**.48**	**.45**	**.56 [.62]**
Anxious Depressed	**.57 [.65]**	**.30** (.00)	**.56**
Withdrawn Depressed	**.41 [.46]**	**.25** (−.01)	**.41**
Rule Breaking	.07 (−.15)	**.40 [.46]**	**.27**
Aggressive	**.35** (−.10)	**.52 [.63]**	**.52**
DSM Affective	**.55 [.55]**	**.34** (−.01)	**.55**
DSM Anxiety	**.64 [.64]**	**.22** (−.03)	**.54**
DSM Oppositional	.18 (−.08)	**.65 [.65]**	**.49**
DSM Conduct	.15 (−.14)	**.53 [.53]**	**.41**

Note. Correlations in bold font are significant at the .01 level (two-tailed). Correlations in parentheses represent discriminant validity coefficients using residualized criteria. Correlations in brackets represent convergent validity coefficients without removing overlapping items on the Youth Self-Report.

pared every convergent validity coefficient with every zero-order discriminant validity coefficient separately for the BPC Internalizing and Externalizing scales. Again, for all comparisons, the convergent coefficients were significantly higher than the discriminant coefficients ($p < .05$). Data from 3 months post intake ($n = 133$) showed results similar to those from the child interviews, with higher discriminant coefficients overall (*r*s ranged from .23 to .29), but in all cases significantly lower than their convergent counterparts.

Cross-Informant Agreement

We examined the agreement of the child and caregiver report from initial administration of BPC ($N = 184$). These correlations were as follows: $r_{\text{Internalizing}} = .22$; $r_{\text{Externalizing}} = .31$; $r_{\text{Total}} = .19$, which were comparable with those found in meta-analytic reviews of parent–child symptom agreement (e.g., Achenbach, McConaughy, & Howell, 1987, reported an average parent–child correlation of .25 and higher cross-informant correlations for externalizing than internalizing across diverse types of cross-informant pairs).

Agreement With Diagnostic Interview

As a final test of validity, we chose to examine how well the BPC Internalizing and Externalizing scales would differ as a function of diagnostic groupings as determined by ChIPS interview data from the separate parent and child interviews. We created two pairs of complementary diagnostic sets: (a) Any Externalizing (any diagnosis of oppositional defiant disorder, conduct

Table 5
Correlations Among Brief Problem Checklist (BPC) Caregiver Scales and Selected Scales From the Child Behavior Checklist (N = 167)

Child Behavior Checklist	BPC CAREGIVER Internalizing	BPC CAREGIVER Externalizing	BPC CAREGIVER Total
Internalizing	**.51 [.57]**	.11 (.00)	**.39**
Externalizing	.18 (.06)	**.62 [.68]**	**.51**
Total	**.33**	**.42**	**.48 [.56]**
Anxious Depressed	**.46 [.54]**	.16 (.00)	**.38**
Withdrawn Depressed	**.37 [.42]**	.09 (−.01)	**.29**
Rule Breaking	.02 (.04)	**.53 [.51]**	**.35**
Aggressive	.19 (.05)	**.58 [.67]**	**.54**
DSM Affective	**.51 [.51]**	**.21** (.01)	**.44**
DSM Anxiety	**.45 [.45]**	−.03 (−.05)	**.26**
DSM Oppositional	.07 (.01)	**.67 [.67]**	**.46**
DSM Conduct	.00 (.03)	**.57 [.57]**	**.35**

Note. Correlations in bold font are significant at the .01 level (two-tailed). Correlations in parentheses represent discriminant validity coefficients using residualized criteria. Correlations in brackets represent convergent validity coefficients without removing overlapping items on the Child Behavior Checklist.

disorder, or disruptive behavior disorder not otherwise specified) versus No Externalizing (children with none of those disorders), and (b) Any Internalizing (any anxiety disorder diagnosis, major depressive or dysthymic disorder diagnosis, or adjustment disorders with depressed or anxious mood) versus No Internalizing (children with none of those disorders). For each scale, we tested whether these groupings (as created by the corresponding informant; e.g., we used child ChIPS interview results to test the child BPC scales) would result in significant between-groups differences on the expected scales. As can be seen in Table 6, these groupings uniformly showed significant differences in the expected direction, with small to medium effect sizes.

Measurement of Change Over Time

We sought to evaluate whether the BPC measure administered weekly provided a within-individual estimate of the slope of change over time that was comparable with the estimates obtained using the longer CBCL and YSR measures. The first step was to obtain reliability estimates from random coefficient growth models for each measure, using a log function of days as the Level 1 predictor and participant (child) as the Level 2 grouping variable. Reliability is generally defined as the ratio of true to observed scores, and in the context of random coefficient models, the reliability of each coefficient may be calculated as the ratio of parameter variance to total variance (parameter variance + error variance). Slope reliabilities are thus a function of (a) the degree to which slopes differ across individuals, and (b) the precision with which each individual's slope is estimated (Raudenbush & Bryk, 2002). The index routinely reported by the hierarchical linear modeling software and shown in Table 7 is an average of the reliability values across all individuals in the sample. Table 7 shows that slope reliability estimates were higher for the BPC–child measure than for the YSR, over both 6- and 12-month periods. Similarly, slope reliability estimates were higher for the BPC–parent measure than for the CBCL, though the differences were not as pronounced as those for the child measures.

We next tested to see whether these random coefficient model-derived slope estimates would correlate across measures—in other words, whether the best estimates of BPC–caregiver and CBCL slopes, and BPC–child and YSR slopes, correlated with one another. Specifically, data from all available BPC interviews within informant for each participant were regressed on log-transformed days in a random coefficients growth model. Individual empirical Bayes slope estimates were obtained for Internalizing, Externalizing, and Total score on all scales over two periods: once using data from baseline to 6 months and once using data from baseline to 12 months. All correlations were statistically significant (see Table 8), with the slopes of the BPC scales moderately correlated with the corresponding slopes on the CBCL and YSR.

Finally, we chose to investigate how the briefer measures with more data collection points would fare relative to the longer measure with fewer data collection points. For this, we examined how well the BPC data would predict CBCL and YSR observed values at 6 months, and we compared this with how well CBCL and YSR scores from baseline and 3 months predicted CBCL and YSR observed scores at 6 months. These correlations appear in Table 9 and show that predicted values using the CBCL and YSR at baseline and 3 months performed similarly to the BPC data in predicting CBCL and YSR at 6 months, with the CBCL performing slightly better than the BPC in the parent analysis, and with the BPC performing slightly better in the child analysis (differences not significant).

Timing and Effort

To determine the degree of effort and the amount of time required to perform the BPC interviews, we gathered data on all calls within a 3-week window for 43 children, representing all participants receiving treatment at that time. This procedure yielded data from 70 child phone interviews and 72 caregiver phone interviews. Any phone call made to obtain responses on a BPC interview was defined as an attempt, and calls that produced responses for all 12 items were considered a success. The mean number of attempts required for successful calls was 2.38 (SD = 2.13) for caregiver interviews and 2.23 (SD = 1.96) for child interviews. These distributions were positively skewed, with the modal number of attempts being a single call for both caregiver and child interviews.

Outcomes of first calls were similar across caregiver and child interviews. For caregivers, first calls yielded a completed assessment 42.3% of the time, involved no answer or leaving a message 48.7% of the time, and involved an informant asking to be called back later 9.0% of the time. For children, first calls yielded a

Table 6
Analysis of Variance Testing Differences Between Internalizing and Externalizing Diagnostic Groups on Child and Caregiver Brief Problem Checklist (BPC) Scales at Intake

Dependent variable	Groups	*n*	*M*	*SD*	*F*	*dfs*	*p*	η^2
BPC Child Internalizing	Any internalizing	68	3.66	2.70	15.34	1, 160	<.001	.09
	No internalizing	94	2.13	2.26				
BPC Child Externalizing	Any externalizing	60	4.28	2.08	195.27	1, 160	<.001	.23
	No externalizing	102	2.01	1.97				
BPC Caregiver Internalizing	Any internalizing	98	5.44	3.01	29.58	1, 160	<.001	.16
	No internalizing	64	2.91	2.72				
BPC Caregiver Externalizing	Any externalizing	102	6.35	2.63	61.62	1, 160	<.001	.28
	No externalizing	60	3.05	2.51				

Table 7
Reliabilities of the Slope of Log Days Predicting BPC and ASEBA Scores

	Period					
	Baseline to 6 months			Baseline to 12 months		
Measure	Internal	External	Total	Internal	External	Total
BPC–child	.85	.85	.88	.86	.86	.88
YSR	.50	.42	.41	.54	.58	.55
BPC–caregiver	.79	.70	.75	.62	.45	.53
CBCL	.82	.73	.78	.67	.55	.57

Note. BPC = Brief Problem Checklist; ASEBA = Achenbach System for Empirically Based Assessment; YSR = Youth Self-Report; CBCL = Child Behavior Checklist.

completed assessment 52.0% of the time, involved no answer or leaving a message 46.7% of the time, and involved an informant asking to be called back later 1.3% of the time.

Callers kept track of the amount of call time required to administer the 12-item checklist during the call. The average time to administer the BPC items was 57 s (*SD* = 13) for caregivers and 53 s (*SD* = 15) for children. On average, then, each BPC interview required less than a minute to administer. The average total call time was 3:48 [minutes:seconds] (*SD* = 3:29) for parents and 2:22 (*SD* = 1:12) for children.

Discussion

The findings are overall quite supportive of the psychometric strength of the BPC interviews in this diverse sample of clinically referred youths with heterogeneous and highly comorbid psychopathology. As expected, both the child interview and caregiver interview yielded two factors that corresponded to the internalizing and externalizing constructs that were intended by the interview design to reflect the structure of the CBCL and YSR. Despite the small number of items on each subscale, internal consistency and test–retest procedures suggested good reliability. Convergent validity was demonstrated through significant correlations between each BPC child interview scale and its corresponding broadband (rs > .50), narrowband (rs > .40), and *DSM*-derived (rs > .50) scales on the YSR, when overlapping items were removed from the YSR. A similar pattern of significant associations was evident between each BPC parent interview scale and corresponding broadband (rs > .50), narrowband (rs > .35), and *DSM*-derived (rs > .50) scales on the CBCL, when overlapping items were removed. Analyses also yielded evidence of discriminant validity, particularly when we controlled for nontarget variance in the criterion measures.

By representing a number of different factors related to measurement, the overall investigation strategy provided an incidental glimpse of the relative contribution of method variance (brief oral instrument vs. long paper instrument), construct (i.e., "true score") variance, time interval variance (Time 1 vs. Time 2), and error variance in the observed data. The cross-method, within-informant, within-construct coefficients ranged from .45 to .56 for the BPC child interview and from .48 to .62 for the BPC parent interview, suggesting that the variance due to change in procedures

Table 8
Correlations Between Empirical Bayesian Estimated Within-Subject Slopes of BPC Interviews and ASEBA Measures Over 6- and 12-Month Time Intervals

Scale	6 months	12 months
	BPC–child	
YSR		
Internalizing	.48	.43
Externalizing	.37	.32
Total	.41	.37
	BPC–caregiver	
CBCL		
Internalizing	.39	.42
Externalizing	.40	.34
Total	.40	.37

Note. All correlations are significant at the .01 level (two-tailed). Slopes are within-subject estimates of scales scores over log function of time in days. BPC = Brief Problem Checklist; ASEBA = Achenbach System for Empirically Based Assessment; YSR = Youth Self-Report; CBCL = Child Behavior Checklist.

Table 9
Correlations Between Observed ASEBA Scores at 6 Months and Predicted ASEBA Scores as a Function of Using Either BPC and ASEBA Data as Predictors

	Predictor	
Criterion	BPC–caregiver	CBCL
CBCL scores observed at 6 months		
Internalizing	.65	.77
Externalizing	.58	.72
Total	.61	.75
	BPC–child	YSR
YSR observed scores at 6 months		
Internalizing	.48	.42
Externalizing	.42	.33
Total	.46	.34

Note. All correlations are significant at the .01 level (two-tailed). ASEBA = Achenbach System for Empirically Based Assessment; BPC = Brief Problem Checklist; CBCL = Child Behavior Checklist; YSR = Youth Self-Report.

(oral administration vs. paper) as well as a substantially reduced item set were relatively minimal (i.e., these ASEBA–BPC validity coefficients approached the ceiling established by their test–retest coefficients).

In terms of the variance due to time interval, as noted, all estimates of convergent and discriminant validity in this study (see Tables 4 and 5) were subject to a separation in time between administrations. Thus, the upper limit of a convergent validity coefficient was not 1.0 but rather the scale's own retest reliability. In general, this means that validity coefficient estimates were underestimated to some degree (especially when removing overlapping BPC items from the CBCL and YSR); convergent validity estimates for simultaneous administration would likely have been higher but so too would have the estimates for discriminant validity coefficients. This is particularly true given that the intertest interval was on average somewhat longer for the validity tests than for the test–retest estimates. For the most precise estimate of validity coefficients, future psychometric research on the BPC interviews might include validity criteria that are gathered concurrently with the BPC to the extent possible.

With respect to effort and timing, our data suggest that these interviews require some modest effort. Although the modal number of calls to complete the BPC was one for both parents and children, the mean was more than two calls. However, on average, calls yielded interview data for roughly every other call attempt. The timing data suggest that the testing burden is quite minimal for client families, with items being administered on average in less than 1 min. These findings were consistent with the overall goal of producing a quick and efficient method, with low participant burden, for gathering frequent measures of clinical progress.

One limitation to the study involves the small number of validity criteria. A stronger design might have incorporated more measures against which to test the BPC. For example, in the longitudinal analyses, it would have been helpful to see whether the BPC interview could predict an independent future criterion as well as the CBCL and YSR scales. Nevertheless, the goals of this first study were to test whether the BPC interviews efficiently provide similar information relative to the longer scales from which they were adapted. The findings are indeed supportive, even when estimates of agreement were based on scales that removed overlapping items from the CBCL and YSR. Further, the BPC interviews show convergence with diagnostic groupings obtained from independent structured clinical interview data. Finally, in the longitudinal tests, the BPC performed comparatively well with the CBCL and YSR measures predicting themselves. An additional limitation involved our inability to examine the performance of this measure within specific age groupings because of the limited sample size. We recommend that future research on the BPC examine its performance as a function of age and that, until then, the measure be used cautiously at the upper and lower age boundaries of our present sample.

Regarding clinical utility, the BPC interviews seem well suited to their particular purpose: creating a brief and standardized means of estimating trajectories on dimensions broad enough to be applicable across a wide variety of problems and contexts. Given the current findings, it appears that estimating trends on such data gathered weekly—in this case, over a 3-month interval or more—has utility for indicating the course of clinical progress. That said, we believe that the BPC interviews should be one component within the context of a larger evidence-based assessment strategy that includes administration of a comprehensive baseline battery at the front end to inform treatment selection and planning, repeated administration of well-established broadband measures (e.g., ASEBA) but on a less frequent schedule (e.g., every 3 or 6 months) to complement and support information from the BPC, and the administration of idiographic progress measures (e.g., specific client goals, narrow-band symptom scales) to index change on client-specific targets. That said, the BPC appears well suited for its primary function of yielding feedback on general outcome metrics on a frequent enough schedule so as to indicate the need for closer review and possible adaptation of the treatment plan.

Conclusion

The present study provides initial support for the psychometric properties of the BPC child and caregiver interviews. In a multiethnic clinical sample of youths, this BPC assessment procedure demonstrated some of the benefits of longer, established measures using a brief, efficient data collection strategy. The present study suggests that the BPC is a promising method for gathering information from youths and their caregivers regarding internalizing and externalizing symptoms, both at a baseline assessment and during the course of clinical care. Given the current findings that the BPC and ASEBA measures appeared to assess related but not identical information, it is recommended that a measurement approach to monitor change using frequent administration of the BPC also include periodic assessments using these longer, well-established measures.

References

Achenbach, T. M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34,* 541–547.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2003). *DSM*-oriented and empirically based approaches to constructing scales from the same item pools. *Journal of Clinical Child and Adolescent Psychology, 32,* 328–340.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22,* 37–53.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Brooks, S. J., & Kutcher, S. (2003). Diagnosis and measurement of anxiety disorder in adolescents: A review of commonly used instruments. *Journal of Child and Adolescent Psychopharmacology, 13,* 351–400.

Chamberlain, P., & Reid, J. B. (1987). Parent observation and report of child symptoms. *Behavioral Assessment, 9,* 97–109.

Chorpita, B. F., Bernstein, A. D., Daleiden, E. L., & the Research Network on Youth Mental Health. (2008). Driving with roadmaps and dashboards: Using information resources to structure the decision models in service organizations. *Administration and Policy in Mental Health and Mental Health Services Research, 35,* 114–123.

Clark, L. A., & Watson, D. B. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fristad, M. A., Cummins, J., Verducci, J. S., Teare, M., Weller, E., & Weller, R. A. (1998). Study IV: Concurrent validity of the *DSM–IV* Revised Children's Interview for Psychiatric Syndromes (ChIPS). *Journal of Child and Adolescent Psychopharmacology, 8,* 227–236.

Fristad, M. A., Glickman, A. R., Verducci, J. S., Teare, M., Weller, E. B., & Weller, R. A. (1998). Study V: Children's Interview for Psychiatric Syndromes (ChIPS): Psychometrics in two community samples. *Journal of Child and Adolescent Psychopharmacology, 8,* 237–245.

Fristad, M. A., Teare, M., Weller, E. B., Weller, R. A., & Salmon, P. (1998). Study III: Development and concurrent validity of the Children's Interview for Psychiatric Syndromes–Parent Version (P-ChIPS). *Journal of Child and Adolescent Psychopharmacology, 8,* 221–226.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3,* 29–51.

Ilardi, S. S., & Craighead, W. E. (1994). The role of nonspecific factors in cognitive-behavior therapy for depression. *Clinical Psychology: Science and Practice, 1,* 138–156.

Klein, D. N., Dougherty, L. R., & Olino, T. M. (2005). Toward guidelines for evidence-based assessment of depression in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34,* 412–432.

Lambert, M., Harmon, C., Slade, K., Whipple, J., & Hawkins, E. (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology, 61,* 165–174.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Reich, W., Shayka, J. J., & Taibleson, C. (1991). *Diagnostic Interview for Children and Adolescents (DICA-R-C): Child Version.* St. Louis, MO: Washington University School of Medicine.

Rooney, M. T., Fristad, M. A., Weller, E. B., & Weller, R. A. (1999). *Administration manual for the ChIPS.* Washington, DC: American Psychiatric Association.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Swenson, L. P., Esposito-Smythers, C., Hunt, J., Hollander, B. L., Dyl, J., Rizzo, C. J., . . . Spirito, A. (2007). Validation of the Children's Interview for Psychiatric Syndromes (ChIPS) with psychiatrically hospitalized adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 46,* 1482–1490.

Teare, M., Fristad, M. A., Weller, E. B., Weller, R. A., & Salmon, P. (1998a). Study I: Development and criterion validity of the Children's Interview for Psychiatric Syndromes (ChIPS). *Journal of Child and Adolescent Psychopharmacology, 8,* 205–211.

Teare, M., Fristad, M. A., Weller, E. B., Weller, R. A., & Salmon, P. (1998b). Study II: Concurrent validity of the *DSM–III–R* Children's Interview for Psychiatric Syndromes (ChIPS). *Journal of Child and Adolescent Psychopharmacology, 8,* 213–219.

Ware, J. E., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation [Special issue]. *Rehabilitation Psychology, 50,* 71–78.

Webster-Stratton, C., & Spitzer, A. (1991). Development, reliability, and validity of the daily telephone discipline interview. *Behavioral Assessment, 13,* 221–239.

Weisz, J. R., Chu, B. C., & Polo, A. J. (2004). Treatment dissemination and evidence-based practice: Strengthening intervention through practitioner–researcher collaboration. *Clinical Psychology: Science and Practice, 11,* 300–307.

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61,* 671–689.

Weisz, J. R., Southam-Gerow, M. A., Gordis, E. B., Connor-Smith, J. K., Chu, B. C., Langer, D. A., . . . Weiss, B. (2009). Cognitive-behavioral therapy versus usual clinical care for youth depression: An initial test of transportability to community clinics and clinicians. *Journal of Consulting and Clinical Psychology, 77,* 383–396.

Weller, E. B., Weller, R. A., Rooney, M. T., & Fristad, M. A. (1999a). *ChIPS: Children's Interview for Psychiatric Syndromes.* Washington, DC: American Psychiatric Association.

Weller, E. B., Weller, R. A., Rooney, M. T., & Fristad, M. A. (1999b). *Children's Interview for Psychiatric Syndromes–Parent Version (P-ChIPS*). Washington, DC: American Psychiatric Association.

Received October 6, 2009
Revision received February 5, 2010
Accepted February 8, 2010 ■