Journal of Clinical Child & Adolescent Psychology

Routledge
Taylor & Francis Group

ISSN: 1537-4416 (Print) 1537-4424 (Online) Journal homepage: http://www.tandfonline.com/loi/hcap20

Efficient Monitoring of Treatment Response during Youth Psychotherapy: The Behavior and Feelings Survey

John R. Weisz, Rachel A. Vaughn-Coaxum, Spencer C. Evans, Kristel Thomassin, Jacqueline Hersh, Erica H. Lee, Mei Yi Ng, Nancy Lau, Jacquelyn N. Raftery-Helmer & Patrick Mair

To cite this article: John R. Weisz, Rachel A. Vaughn-Coaxum, Spencer C. Evans, Kristel Thomassin, Jacqueline Hersh, Erica H. Lee, Mei Yi Ng, Nancy Lau, Jacquelyn N. Raftery-Helmer & Patrick Mair (2019): Efficient Monitoring of Treatment Response during Youth Psychotherapy: The Behavior and Feelings Survey, Journal of Clinical Child & Adolescent Psychology

To link to this article: https://doi.org/10.1080/15374416.2018.1547973

Journal of Clinical Child & Adolescent Psychology, 00(00), 1–15, 2019
Copyright © Society of Clinical Child & Adolescent Psychology
ISSN: 1537-4416 print/1537-4424 online
DOI: https://doi.org/10.1080/15374416.2018.1547973

Routledge
Taylor & Francis Group

Efficient Monitoring of Treatment Response during Youth Psychotherapy: The Behavior and Feelings Survey

John R. Weisz, Rachel A. Vaughn-Coaxum, and Spencer C. Evans
Department of Psychology, Harvard University

Kristel Thomassin
Department of Psychology, University of Guelph

Jacqueline Hersh
Department of Psychology, Appalachian State University

Erica H. Lee
Department of Psychiatry, Boston Children's Hospital, Harvard Medical School

Mei Yi Ng
Department of Psychology, Florida International University

Nancy Lau
Department of Pediatrics, University of Washington School of Medicine

Jacquelyn N. Raftery-Helmer
Department of Psychology, Worcester State University

Patrick Mair
Department of Psychology, Harvard University

An emerging trend in youth psychotherapy is measurement-based care (MBC): treatment guided by frequent measurement of client response, with ongoing feedback to the treating clinician. MBC is especially needed for treatment that addresses internalizing and externalizing problems, which are common among treatment-seeking youths. A very brief measure is needed, for frequent administration, generating both youth- and caregiver-reports, meeting psychometric standards, and available at no cost. We developed such a measure to monitor youth response during psychotherapy for internalizing and externalizing problems. Across 4 studies, we used ethnically diverse, clinically relevant samples of caregivers and youths ages 7–15 to develop and test the Behavior and Feelings Survey (BFS). In Study 1, candidate items identified by outpatient youths and their caregivers were examined via an MTurk survey, with item response theory methods used to eliminate misfitting items. Studies 2–4 used separate clinical samples of youths and their caregivers to finalize the 12-item BFS (6 internalizing and 6 externalizing items), examine its psychometric properties, and assess its performance in

Correspondence should be addressed to John Weisz, Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138. E-mail: john_weisz@harvard.edu

monitoring progress during psychotherapy. The BFS showed robust factor structure, internal consistency, test–retest reliability, convergent and discriminant validity in relation to three well-established symptom measures, and slopes of change indicating efficacy in monitoring treatment progress during therapy. The BFS is a brief, free youth- and caregiver-report measure of internalizing and externalizing problems, with psychometric evidence supporting its use for MBC in clinical and research contexts.

An emerging form of evidence-based practice in psychotherapy is measurement-based care (MBC): using data on client treatment response, collected throughout episodes of care, to guide intervention (Fortney et al., 2017; Scott & Lewis, 2017). The data can be used by the treating clinician to inform judgments throughout treatment, potentially strengthening intervention. Monitoring client response throughout treatment can tell the clinician how the intervention is working, which treatment foci are and are not showing improvement, and thus when changes in strategy are needed. The feedback can also help clinicians know when treatment goals have been met—important where conditions (e.g., waitlists) require efficiency. MBC data can be used in clinical research—synthesized across sessions to show trajectories of change for individual clients, for subgroups of interest, or for differing treatment conditions—to convey the slope of change and the ultimate outcome (e.g., Chorpita et al., 2013, Chorpita et al., 2017; Weisz, Bearman, Santucci, & Jensen-Doss, 2017, Weisz et al., 2012). Beneficial effects of MBC on client outcomes have been shown in studies and meta-analyses spanning a broad range of client ages and treated problems (e.g., Bickman, Douglas Kelley, Breda, de Andrade, & Riemer, 2011; Fortney et al., 2017; Shimokawa, Lambert, & Smart, 2010).

MBC is increasingly emphasized in psychotherapy with children and adolescents (herein “youths”), in which frequent feedback from both youths and caregivers provides guidance to treating clinicians (De Los Reyes, Augenstein, & Aldao, 2017; De Los Reyes et al., 2015). Given the array of different youth treatments, each distinctive in approach, goals, and target problems, no single measure could adequately address or be appropriate for all treatments of all problems. What is needed, instead, is an array of clinically sensitive measurement options, each designed to fit particular therapy objectives and foci. The focus of the present article is on the measurement needed for MBC with youth therapies addressing the two most robustly identified dimensions of psychopathology in clinically referred youths: internalizing and externalizing problems (e.g., Achenbach, 1966; Achenbach, Conners, Quay, Verhulst, & Howell, 1989; Cicchetti & Toth, 1991; Mash & Wolfe, 2013; Quay, 1979).

Such measurement might be used in everyday practice by clinicians whose caseloads include youths with problems on both dimensions—a common situation in clinical practice settings (see, e.g., Staller, 2006), consistent with the well-documented co-occurrence of these two dimensions (e.g., Achenbach et al., 1989; Youngstrom, Findling, & Calabrese, 2003). Moreover, even youths who are being treated mainly for internalizing or externalizing problems typically have additional problems of the other type that are relevant to treatment and warrant tracking during treatment. Measurement of both internalizing and externalizing problems is also very useful for clinicians using the recent generation of transdiagnostic youth treatments designed to address internalizing and externalizing problems by combining empirically supported treatment components for both forms of dysfunction (e.g., Chorpita & Weisz, 2009; Dorsey, Berliner, Lyon, Pullmann, & Murray, 2014; Weisz et al., 2017). During intervention for internalizing and externalizing problems, a common challenge is the clinical decision making required throughout treatment, as the clinician tries to gauge the young client’s treatment response to determine which problem focus is most appropriate and tailor intervention accordingly (Ng & Weisz, 2016). MBC can help address this challenge by informing clinician judgments, but MBC for treatment of internalizing and externalizing problems is likely to work best when focused specifically on measurement of those two dimensions.

Several features could make a measure of internalizing and externalizing especially useful for MBC purposes. Ideally, such a measure would (a) provide feedback to the treating clinician from both youth and caregiver perspectives; (b) be very brief, minimizing measurement burden so as to keep youths and caregivers engaged and responding throughout treatment; (c) meet accepted psychometric standards and be sensitive to change during treatment; and (d) be free to all, thus eliminating financial barriers to use by clinicians and researchers. Important to note, the need for brevity and precise focus in MBC would mean that not all clinical problems could be included in a single measure. Thus, MBC using standardized assessment, to ensure a common metric for both youth and caregiver and across all clients, can be helpfully complemented by idiographic tracking of the severity of specific problems that each youth and each caregiver identifies as especially important to them (see, e.g., Weisz et al., 2011). The blend of nomothetic, same playing field assessment, with idiographic, completely personalized assessment, may provide a particularly sensitive and clinically useful blend for MBC. An essential partner in this formula is a standardized measure that has the characteristics outlined at the beginning of this paragraph to provide for repeated measurement of the internalizing and externalizing dimensions targeted in treatment and for comparison of change trajectories by parents and youths, and across multiple clients, on the same set of items.

One published measure of internalizing and externalizing problems did have the necessary characteristics, and it was used to good effect for MBC assessment, often in combination with idiographic measurement. This 12-item measure—the Brief Problem Checklist (BPC; Chorpita et al., 2010)—included six internalizing items (encompassing anxiety and depression) and six externalizing items (behavioral/conduct problems). The BPC, administered weekly, performed well—as a guide to clinicians and as a measure of outcome trajectories—within multiple studies of internalizing and externalizing treatment (e.g., Chorpita et al., 2017; Weisz et al., 2017, 2012). However, its items were derived from the Child Behavior Checklist (CBCL) and Youth Self-Report (YSR; Achenbach & Rescorla, 2001), and copyright issues have ruled out further use of the BPC. We searched for available alternative measures for MBC that might have the characteristics specified in the preceding paragraph. We found shortened forms of measures such as the CBCL, YSR, and Youth Outcome Questionnaire (e.g., Burlingame et al., 2001), but their breadth of item content was not a good fit, and their cost would rule out the frequent use required for MBC by clinicians and researchers with modest budgets. A measure called the Symptoms and Functioning Severity Scale has short forms (e.g., Gross, Hurley, Lambert, Epstein, & Stevens, 2015) and is free, but the content is not a very good fit (e.g., items include inattention, hyperactivity, alcohol and drug use), and there is limited evidence on psychometrics of the short forms or how they might function as progress monitoring measures. We considered the Strengths and Difficulties Questionnaire (Goodman, 2001), which is well studied and free; but the full measure has 25 items, most not focused on internalizing or externalizing, and requirements for use prohibit administering the measure more often than monthly (R. Goodman, personal communication, August 21, 2013). Our review of these and other measures identified certain strengths but failed to identify an existing measure of internalizing and externalizing with the characteristics needed for MBC. So we set out to develop such a measure, one that could meet clinician and researcher needs and be free to all users, with no copyright restrictions.

Accordingly, we used a series of studies to develop and evaluate a very brief youth- and caregiver-report measure exclusively focused on internalizing and externalizing problems—a measure that, if psychometrically sound, could be used to monitor client response frequently (e.g., weekly) during treatment of internalizing and externalizing problems. To develop this measure, to be called the Behavior and Feelings Survey (BFS), we developed a pool of candidate items using lists of "top problems" generated by clinically referred youths and their caregivers. Then we carried out a series of four studies that entailed (a) eliminating misfitting items, using data from an online survey (Study 1); (b) using ratings from young mental health clinic clients and their caregivers to finalize the BFS (Study 2); (c) assessing BFS psychometrics using a second clinical sample (Study 3); and (d) using a third clinical sample to assess sensitivity of the BFS to change during treatment, and thus its utility for monitoring progress during therapy (Study 4). For all participants in each of the studies, consent, assent, and all other human subjects procedures were reviewed and approved by the relevant Institutional Review Boards.

STUDY 1: REDUCING AN INITIAL ITEM POOL VIA MTURK

In Study 1, we began with a pool of candidate internalizing and externalizing items and carried out an initial step of the scale reduction needed for an eventual brief measure. Candidate items were derived from "top problems" identified via assessments with clinically referred youths and their caregivers at the outset of therapy. The candidate items were administered to an online sample of parents via Amazon's Mechanical Turk (MTurk; https://www.mturk.com/). Responses were used in an initial step of scale reduction, with exploratory factor analysis (EFA) and item response theory (IRT) analyses used to identify items to exclude from the pool.

Study 1: Method

The initial pool of candidate items was derived from a sample of 178 youths and their caregivers (using the data set from Weisz et al., 2011), seeking mental health care in one of 10 community outpatient programs in office- and school-based settings of two large metropolitan areas. The youths (ages 7–14) and caregivers, interviewed separately just prior to treatment, were asked to identify the three problems they were concerned about and saw as most important to work on in therapy. Clinical representativeness and comprehensiveness were supported by including both youths and caregivers, and from multiple mental health service programs, both clinic based and school based. Relevance to intended measure content was supported by including youths referred for multiple problems, largely internalizing and externalizing, and synthesizing data on the top problems they and their caregivers identified. The process generated 470 youth-identified problems and 514 caregiver-identified problems. We then eliminated (a) nonpsychopathology content (e.g., irritated by sister, strict parents); (b) problems that did not fit internalizing or externalizing dimensions according to youth psychopathology reviews, factor analyses, or meta-analyses (e.g., Achenbach, 1966; Cicchetti & Toth, 1991; Quay, 1979); and (c) problems that did not appear on both youth and caregiver lists. From the remaining content, we sought items with breadth and generality, because highly specific items (e.g., afraid of spiders, fights with Kevin) would fit too few youths to be useful in MBC. Thus, although we retained exact wording of participants for some items, for others we used broader terminology to encompass multiple very specific problems identified by participants; for example, separate top problems identifying specific fears (e.g., the ocean, the dark, getting shots) were encompassed within "Feeling nervous or afraid." We retained partially redundant items initially, so that data analyses could help us determine which wording provided the most psychometrically sound item set.

This process produced 48 items—28 internalizing and 20 externalizing. The internalizing items included depression-related content (e.g., feeling sad, not enjoying things) and anxiety-related content (e.g., feeling nervous or afraid, thinking scary thoughts). The externalizing items included a range of behavioral/conduct problems (e.g., disobeying, telling lies, fighting). Next we analyzed the data to eliminate items that were a poor fit psychometrically.

MTurk Survey and Initial Item Pool Reduction

For this purpose, we used data collected online to assess interitem associations in our collection of 48 items. IRT methods were used to identify the subsets of items that taken together could capture the latent traits of internalizing and externalizing and to eliminate candidate items that did not fit. This was considered a preliminary step, in part because MTurk can be used to obtain adult responses only (i.e., of parents but not their children). The goal was to shrink the item pool to a manageable list that appeared psychometrically appropriate based on interitem associations derived from parent reports, with further scale reduction, reliability, and validity testing to be carried out in Studies 2–4 (presented later).

Parents' ratings of their children's problems on the 48 items were generated via MTurk. Here and in Studies 2–4, items were rated on a scale from 0 (*not a problem*) to 4 (*a very big problem*), with no descriptive anchors provided for 1, 2, 3, to emphasize the continuous interval properties of the scale. The sample included 585 adults with valid responses, self-identified as living in the United States and speaking English fluently, and as parents of at least one youth between age 7 and 15 who (a) had received, was currently receiving, or could benefit from counseling or mental health services or (b) had moderate to severe emotional or behavioral problems. With MTurk registration limited to adults, our sample included only parents. To be included, parents had to correctly answer three of the four attention test questions embedded in the survey; parents were paid $2 for their participation, an amount within the midrange of MTurk payment. Table 1 shows sample characteristics. On our family income question, 33% reported an annual income of $0–$39,999, 43% reported $40,000–$79,999, 19% reported $80,000–$119,999, and 5% reported $120,000 or higher. Participants were randomly assigned to rate their child's behavior either during the past month (Subsample M; $N = 282$) or the past week (Subsample W; $N = 303$), permitting us to check the robustness of findings across differing reporting time frames that might be used to monitor treatment response.

Data Analytic Plan

To examine the underlying structure of the BFS items, a maximum-likelihood (ML) EFA was implemented in the R environment for statistical computing (R Core Team, 2015) using the "psych" package (Revelle, 2018). The best fitting factor solution was identified via visual examination of the scree plot and an established goodness of fit indicator, the Very Simple Structure (VSS) criterion, which has been shown (Revelle & Rocklin, 1979) to more consistently identify the optimal number of factors to extract when compared to ML tests and Kaiser's eigenvalue cutoff of 1.0 (Kaiser, 1960). To examine and reduce the 48-item set, each factor extracted was further examined under an IRT approach, using Rating Scale Models (RSM). IRT is a theory of estimation, wherein estimates of individuals' latent trait scores are derived from the latent properties of both the scale items and the individuals and are estimated independently of the sample characteristics from which they were derived (de Ayala, 2009).

Each factor extracted from the EFA was fit to a RSM using the "eRm" package in R (Mair & Hatzinger, 2007), as items submitted to IRT models must reflect unidimensional constructs. When validating scales comprise multiple factors, each factor or singular dimension must be modeled separately. Item fit was examined via chi-square tests, and items with significant chi-square values were eliminated. The RSM was refit to the data, iteratively, each time an item was eliminated to ensure that removal of a single item did not change the fit of remaining items within the set. The process was repeated until all items remaining showed good fit. Mean-square infit and outfit statistics were also examined as complementary indices of item fit. Infit and outfit statistics less than 2.0 suggest acceptable fit, whereas statistics at 1.5 or less are ideal and indicate that the item is productive for measurement and not likely redundant. A likelihood-ratio test was used to determine whether items showed differential functioning when participants were instructed to report on symptoms over 1 month (M) versus 1 week (W). Nonsignificant *p* values would indicate equivalent functioning of items between Subsample M and Subsample W.

Study 1: Results and Discussion

Results from the ML EFA supported one general factor and a two-factor solution with the 28 items originally identified as internalizing loading distinctly onto a single factor (loadings = 0.25–0.84) and the 20 externalizing items loading distinctly onto a second factor (loadings = 0.56–0.90). The two-factor solution accounted for 50% of the variance. Examination of the scree plot was consistent, with only these two factors falling above the "elbow" criterion. The VSS achieved a maximum value of 0.79 with two factors extracted. Internal

TABLE 1
Sample Characteristics for Studies 1 to 4

	Study 1	*Study 2*	*Study 3*	*Study 4*
N[a]	585	222	79	95
Age, *M* (*SD*)	10.8 (2.6)	10.9 (2.3)	10.8 (2.4)	9.8 (1.7)
Age Range	7–15	7–15	7–15	7–14
Sex: Female, %	37.4	48.2	50.0	40.4
Ethnicity, %				
Black/African American	7.9	19.5	25.3	11.7
Latino/Hispanic	6.0	16.7	26.6	4.3
White/Caucasian	76.8	47.1	32.9	54.3
Asian/Asian American	0.9	1.4	1.3	6.4
Multiracial	7.2	14.0	11.4	20.2
Other	1.4	1.4	2.5	3.2
Caregiver Reporters, %				
Mothers	69.2	87.1	91.0	83.2
Fathers	30.8	6.4	6.6	8.4
Other Caregivers	0.0	6.5	2.4	8.4
Baseline Raw Scores, *M* (*SD*)[b]				
BFS Caregiver Internalizing	—	6.31 (6.11)	7.66 (6.68)	8.40 (5.52)
BFS Caregiver Externalizing	—	10.15 (7.56)	12.20 (7.86)	10.75 (7.15)
BFS Caregiver Total	—	16.46 (10.02)	19.86 (9.91)	19.15 (9.42)
BFS Youth Internalizing	—	5.98 (6.70)	7.49 (6.59)	6.62 (5.46)
BFS Youth Externalizing	—	6.08 (6.12)	7.99 (6.90)	5.24 (5.87)
BFS Youth Total	—	12.05 (10.35)	15.48 (10.59)	11.86 (9.08)
Baseline *T* Scores, *M* (*SD*)[b]				
CBCL Internalizing	—	62.73 (9.86)	65.32 (8.68)	67.32 (6.99)
CBCL Externalizing	—	60.67 (10.30)	63.61 (10.08)	61.13 (9.66)
CBCL Total Problems	—	63.12 (9.53)	66.06 (6.96)	66.11 (7.01)
YSR Internalizing	—	55.46 (12.50)	56.62 (10.45)	57.77 (10.06)
YSR Externalizing	—	51.38 (10.54)	53.78 (10.51)	50.26 (8.68)
YSR Total Problems	—	54.92 (11.86)	56.86 (10.38)	56.68 (9.36)

Note: Percentages for mothers and fathers include mother figures (e.g., stepmothers) and father figures (e.g., boyfriends living in the household).

[a] Study 1 included caregiver-report only, Study 2 had youth-report for 203 cases, and Studies 3 and 4 had both raters for all participants.

[b] In Study 1, Child Behavior Checklist (CBCL) and Youth Self-Report (YSR) were not collected, and the Behavior and Feelings Survey (BFS) data are not reported because the measure was still considered experimental at that point (e.g., with 48 items, different versions of the instructions, and MTurk convenience sample of parents but no youths), and therefore those results may not represent the final BFS or generalize to other samples in the same way as the samples for Studies 2–4.

consistency reliability was high for both the internalizing factor (α = 0.95) and the externalizing factor (α = 0.97).

IRT analyses were performed on the 28 internalizing and 20 externalizing items separately. The results produced 29 items remaining after stepwise elimination: 16 internalizing items and 13 externalizing items with infit and outfit statistics less than 1.5, with all chi-square values nonsignificant, indicating no item misfit. Finally, the results of the likelihood-ratio test were nonsignificant, and visual examination of a goodness-of-fit plot further supported equivalent functioning of the items with the 1-month and 1-week time frames.

To summarize, in Study 1 "top problems" identified by clinically referred youths and their caregivers generated a list of 48 candidate BFS items, and IRT methods were used to eliminate poorly fitting items. The resulting 29-item version was the starting point for Study 2.

STUDY 2: USING A CLINICAL SAMPLE TO FINALIZE THE BEHAVIOR AND FEELINGS SURVEY

In Study 2, we sought to reduce the 29 items that met acceptability criteria in the MTurk sample of Study 1 to a smaller, psychometrically sound item set that would be concise enough for repeated use throughout treatment. Also, because MTurk surveys can be used with adults only, we needed to broaden the sample to include youths and identify an item set that would (a) be common to youths and their caregivers and (b) meet psychometric standards for both groups. In addition, we collected youth and caregiver reports on three widely used and well-studied checklist measures to provide evidence of convergent and discriminant validity. To pursue these objectives, and to support clinical relevance of the ultimate item set, we collected data directly from a clinical sample of youths and their caregivers in multiple outpatient treatment sites.

Study 2 Method

Study 2 included 203 youths (222 asked, 19 declined) and their caregivers ($N = 222$) seeking outpatient treatment in one of 14 community mental health clinics. Table 1 shows sample characteristics. Youths and caregivers, interviewed separately, completed all measures with a trained staff interviewer at the time of clinic intake (T1). There was a second assessment at a target interval of 7 days after T1 (actual lag $M = 6.50$ days, $SD = 4.02$), to reevaluate the factor structure and to examine test–retest reliability. This second assessment (T2) was completed by 78% of caregivers ($N = 174$) and 76% of youths ($N = 154$). Measures administered included the youth and caregiver forms of the 29 candidate BFS items and the following measures.

Child Behavior Checklist and Youth Self-Report (YSR; Achenbach & Rescorla, 2001)

These are parallel 118-item caregiver- and youth-report measures of youths' emotional and behavioral problems. Each item is rated on a 3-point scale of 0 (*not true*), 1 (*somewhat or sometimes true*), and 2 (*very true or often true*). Both measures produce a Total Problems score, two broadband Internalizing and Externalizing syndrome scale scores, plus additional narrowband syndrome scales and *Diagnostic and Statistical Manual of Mental Disorders* diagnosis-related scales (American Psychiatric Association, 2000)—four of which are relevant to the internalizing or externalizing dimensions of the present study: Affective Problems, Anxiety, Oppositional, and Conduct Problems. Researchers have documented internal consistency, reliability, and validity of the CBCL and YSR scales (Achenbach & Rescorla, 2001).

Strengths and Difficulties Questionnaire

The Strengths and Difficulties Questionnaire (SDQ; Goodman, 2001) is a 25-item measure with parallel caregiver- and youth-report forms, assessing a range of emotional and behavioral problems in youths. Each item is rated on a 3-point scale of 0 (*not true*), 1 (*somewhat true*), and 2 (*certainly true*). The five subscales include two that correspond to internalizing and externalizing—that is, emotional problems (e.g., "I am often unhappy, depressed or tearful") and conduct problems (e.g., "I get very angry and often lose my temper")—plus peer problems, inattention/hyperactivity, and prosocial behaviors. Studies have documented internal consistency, reliability, and validity of the subscales and the full summary score of the parent- and youth-report versions (Goodman, 2001; Vostanis, 2006; Wolpert, Cheng, & Deighton, 2015).

Youth Outcome Questionnaire 2.01

The Youth Outcome Questionnaire (YOQ; Burlingame et al., 2001) is a 64-item parent- and youth-report measure assessing a range of youth behavioral and emotional problems. Items are rated on a 5-point scale ranging from *never or almost never* to *almost always or always*. Evidence supports the reliability and validity of the total and subscale scores via both parent and youth report (e.g., Burlingame et al., 2001; Dunn, Burlingame, Walbridge, Smith, & Crum, 2005).

Data Analytic Plan

The factor analytic and IRT procedures of Study 1 were replicated in Study 2 to examine and further reduce the set of 29 items. To maximize reliability, IRT models were fit to the data at T1 and T2 to identify the best fitting items that overlapped across caregiver and youth reports and both assessments. Items that survived the IRT stepwise elimination were subsequently analyzed via two methods. Internal consistency (Cronbach's alpha, mean interitem correlation) was calculated for the internalizing, externalizing, and total problem item sets, and test–retest reliability was calculated for each BFS scale at T1 and T2. Correlations were also calculated between youth and caregiver BFS internalizing, externalizing, and total scales, and convergent and discriminant validity were assessed in relation to relevant scales of the YSR/CBCL, SDQ, and YOQ, with z tests used following Meng, Rosenthal, and Rubin (1992). Little's (2002) MCAR test was nonsignificant, suggesting T2 data were missing completely at random. The test-retest and T2 models used only cases with data at both occasions.

Study 2 Results and Discussion

EFA, Rating Scale Model, and Item Fit Analyses

The results of the ML EFA with a promax rotation supported a two-factor solution for both the caregiver-report and youth-report scales. For the caregiver version of the scale, the 16 internalizing items loaded distinctly onto a single factor (loadings = 0.49–0.83), and the 13 externalizing items loaded distinctly onto a second factor (loadings = 0.67–0.88). No internalizing items loaded onto the externalizing factor (all loadings ≤ 0.18) and no externalizing items loaded onto the internalizing factor (all loadings ≤ 0.10). The two-factor solution accounted for 55.5% of the variance, and the VSS achieved a maximum value of 0.91 with two factors extracted. Comparable results emerged for the youth version of the scale, with the 16 internalizing items loaded distinctly onto a single factor (loadings = 0.48–0.87) with no overlap on the externalizing factor (all loadings ≤ 0.20). The 13 externalizing items loaded distinctly onto a second factor (loadings = 0.65—0.83), with no overlap on the internalizing factor (all loadings ≤ 0.12). The two-factor solution accounted for 51.2% of the variance, and the VSS achieved a maximum value of 0.81 with two factors extracted. An RSM was fit to the youth and caregiver versions of the internalizing and externalizing factors at T1 and T2, resulting in four RSMs. After stepwise elimination of misfitting items, and removal of items showing inconsistent fit across reporters and time points (e.g., items that fit well for youths but not caregivers, or at T1 but not T2), six items remained for the internalizing scale and six items remained for the externalizing scale. Details are provided in the following two paragraphs.

Internalizing Factor

For the T1, parent-report version of the scale, 10 items initially showed good fit after stepwise elimination (MSQ infit = 0.76– 1.05; MSQ outfit = 0.68–1.06). The youth-report version showed similarly good fit for 11 items (MSQ infit = 0.79–1.12; MSQ outfit = 0.49–1.18). For the T2 parent-report version of the scale, nine items initially showed good fit after stepwise elimination (MSQ infit = 0.73–1.09; MSQ outfit = 0.76–1.18). The youth-report version of the scale showed similarly good fit for nine items (MSQ infit = 0.77–1.13; MSQ outfit = 0.62–1.07). Across the reduced sets of best-fitting items for all reporters and time points, six items (see Table 2) overlapped across parent and youth report and T1 and T2 assessments. All four models were then refit with the final six items, demonstrating good fit in each case. Although the criterion for item elimination was set at $p < .002$ for the chi-square test, all items produced nonsignificant results except for one item at T2 on the youth-report version of the scale. The item "I think sad or scary thoughts over and over again", produced a significant chi-square test at T2 (but not at T1) for youth but not caregiver report ($p = .023$). This item was retained, however, as all infit and outfit statistics were in the optimal range.

Externalizing Factor

For the T1 caregiver-report version of the scale, eight items initially showed good fit after stepwise elimination (MSQ infit = 0.79–1.09; MSQ outfit = 0.77–-1.03). The youth-report version of the scale showed similarly good fit for nine items (MSQ infit = 0.80–1.04; MSQ outfit = 0.76–1.02). For the T2 caregiver-report version of the scale, 10 items initially showed good fit after stepwise elimination (MSQ infit = 0.68–1.08; MSQ outfit = 0.67–1.18). The youth-report version of the scale showed similarly good fit for nine items (MSQ infit = 0.70–1.24; MSQ outfit = 0.64–1.18). Across the reduced sets of best-fitting items across reporters and time points, six externalizing items (see Table 2) overlapped across caregiver and youth report and T1 and T2. Each of the four models was then refit with the final six items, demonstrating good fit in each case (MSQ infit = 0.73–1.24; MSQ outfit = 0.64–1.22). Youth- and caregiver-report versions of the final set of 12 items are shown in Table 2.

Internal Consistency, Test–Retest Reliability, and Scale Correlations

Next, we assessed internal consistency (Cronbach's alpha, *M* interitem correlations) and test–retest reliability of the full 12-item BFS and the six-item internalizing and externalizing scales per caregiver and youth report for T1 and T2. All estimates are reported in the lower portion of Table 2. As shown, internal consistency was generally good to excellent across all time points, informants, and subscales (αs = .85–.94; *M* interitem *r*s = .31–.73). Similarly, T1–T2 test–retest reliability was consistently high (*r*s = .66–.79). Youth-reported internalizing and externalizing scales were moderately correlated with one another ($r = .30$, $p < .001$), whereas both were more strongly correlated with BFS total problems (internalizing-total $r = .83$, $p < .001$; externalizing-total $r = .79$, $p < .001$). In contrast, caregiver-reported internalizing and externalizing were weakly correlated with each other ($r = .06$, *ns*), but each of these scales was strongly correlated with BFS total problems (internalizing-total $r = .66$, $p < .001$; externalizing-total $r = .79$, $p < .001$).

Convergent and Discriminant Validity

Convergent and discriminant validity analyses of T1 data focused on the relation between BFS scores and other youth psychopathology measures collected at baseline. The findings, presented in Table 3, show the correlations between BFS youth- and caregiver-report internalizing and externalizing scale scores and other measures' scales intended to correspond (i.e., shown in bold) or differ (i.e., discriminant coefficients, not in bold) in internalizing versus externalizing content. Discriminant coefficient *z*-test results are shown in the fourth column; all these *z* values were significant, nearly all at less than .001, indicating strong discriminant validity of the BFS internalizing and externalizing scales. BFS caregiver-report total problems score was highly correlated with CBCL total problems ($r = .67$, $p < .001$), YOQ-caregiver total ($r = .78$, $p < .001$), and SDQ-caregiver total ($r = .56$, $p < .001$). Similarly, BFS youth-report total problems score was highly correlated with YSR total problems ($r = .73$, $p < .001$) YOQ-youth total ($r = .78$, $p < .001$), and SDQ-youth total ($r = .64$, $p < .001$).

To summarize, in Study 2 we applied EFA and fit statistics to data from a clinical sample of youths and caregivers to reduce the 29-item set to a 12-item BFS suitable for frequent use throughout treatment. Responses by the same sample to three well-established checklist measures permitted assessment of convergent and discriminant validity of the 12-item BFS, setting the stage for additional analyses in Study 3.

TABLE 2
Study 2 Promax-Rotated Factor Structure, Internal Consistency, and Reliability Estimates for Final 12-Item Behavior and Feelings Survey (BFS) at Time 1 and Time 2

	Time 1				*Time 2*			
	Caregiver[a]		*Youth*[b]		*Caregiver*[c]		*Youth*[d]	
BFS Scale Items and Factor Loadings	*F1*	*F2*	*F1*	*F2*	*F1*	*F2*	*F1*	*F2*
Internalizing								
Feel Sad	.06	**.81**	−.08	**.85**	.04	**.76**	.06	**.71**
Feel Bad About/Don't Like Self	.03	**.65**	.06	**.71**	.00	**.79**	−.12	**.64**
Feel Down or Depressed	.10	**.84**	−.01	**.81**	.05	**.84**	−.04	**.84**
Feel Nervous or Afraid	−.05	**.72**	.01	**.86**	−.04	**.79**	−.09	**.86**
Worry Bad Things Happening	−.05	**.77**	.04	**.75**	−.03	**.72**	.12	**.75**
Think Sad/Scary Thoughts	−.05	**.69**	−.01	**.71**	−.03	**.69**	.12	**.59**
Externalizing								
Talk Back/Argue Parents/Adults	**.85**	.04	**.82**	−.06	**.86**	.06	**.82**	−.06
Refuse to Do What You Are Told	**.84**	.01	**.74**	−.04	**.81**	.06	**.74**	.03
Do Things Not Supposed to Do	**.76**	−.05	**.77**	.12	**.80**	−.09	**.77**	.08
Rude or Disrespectful	**.90**	.05	**.73**	.06	**.79**	.06	**.73**	.06
Argue With People	**.89**	.04	**.84**	−.03	**.83**	.06	**.84**	−.07
Break Rules at Home/School	**.80**	−.06	**.78**	−.06	**.89**	−.10	**.78**	−.03

Internal Consistency and Reliability	*T1 Caregiver*	*T1 Youth*	*T2 Caregiver*	*T2 Youth*
Internalizing				
Cronbach's α	.88	.90	.89	.87
M Interitem *r*	.56	.61	.59	.53
Test–Retest *r*	—	—	.66	.76
Externalizing				
Cronbach's α	.94	.90	.93	.89
M Interitem *r*	.73	.61	.69	.59
Test–Retest *r*	—	—	.79	.77
Total				
Cronbach's α	.85	.88	.88	.88
M Interitem *r*	.31	.39	.38	.39
Test–Retest *r*	–	–	.78	.78

Note: Two factors were extracted from the Promax-rotated, ordinal exploratory factor analysis, providing evidence of distinct internalizing and externalizing factors. Factor loadings for each item on each of the two factors are presented by reporter (caregiver, youth) and occasion (Time 1 [T1], Time 2 [T2]). Externalizing items consistently loaded most strongly on Factor 1 (F1), across reporter and across assessment time points. Internalizing items consistently loaded most strongly on Factor 2 (F2), across reporters and across assessment time points. Estimates of reliability and internal consistency were consistently good across caregiver and youth reports for Internalizing and Externalizing subscales, at the two time points. Youths and caregivers rated the 12 problems for severity during the past week; ratings could range from 0 (*not a problem*) to 4 (*a very big problem*).

[a] *N* = 222.

[b] *N* = 203.

[c] *N* = 174.

[d] *N* = 154.

STUDY 3: PSYCHOMETRIC ASSESSMENT OF THE 12-ITEM BFS ADMINISTERED IN ITS FINAL FORM

The results from the EFA and IRT analyses of Study 2 led to reduction of the 29-item set to 12 items, six internalizing and six externalizing, with the psychometric characteristics needed for the final BFS. The 12-item BFS showed appropriate internal consistency, test–retest reliability, and convergent and discriminant validity internally and relative to prominent standardized caregiver- and youth-report problem measures. Because Study 2 investigated the BFS within the mix of 29 total items, we carried out Study 3 to assess how the 12-item version would perform when no additional items were included, and in relation to the widely used CBCL and YSR. Thus, Study 3 provided a test of whether the psychometric findings of Study 2 would replicate in a new clinical sample when the final version of the BFS was used.

TABLE 3
Study 2 Correlations of 12-Item BFS Scales for Youth- and Caregiver-Report With Other Youth- and Caregiver-Report Problem Measures

Youth Report	*BFS Youth Internalizing*	*BFS Youth Externalizing*	*Discriminant Z Score*
YSR[a]			
Internalizing	**.73*****	.35***	5.25***
Externalizing	.31***	**.74*****	−6.28***
DSM Affective Problems	**.69*****	.29***	5.48***
DSM Anxiety	**.61*****	.32***	3.76***
DSM Oppositional	.20**	**.69*****	−6.44***
DSM Conduct Problems	.24***	**.69*****	−6.02***
YOQ[b]			
Intrapersonal Distress	**.77*****	.39***	5.18***
Interpersonal Relations	.43***	**.65*****	−2.69***
Social Problems	.11	**.52*****	−3.97***
Behavioral Dysfunction	.44***	**.63*****	−2.23*
Somatic	**.56*****	.23**	3.39***
SDQ[c]			
Emotional Symptoms	**.74*****	.23**	6.14***
Conduct Problems	.20*	**.66*****	−5.06***
Caregiver Report	*BFS Caregiver Internalizing*	*BFS Caregiver Externalizing*	*Discriminant Z Score*
CBCL[d]			
Internalizing	**.72*****	.12	8.24***
Externalizing	.06	**.77*****	−10.05***
DSM Affective Problems	**.66*****	.11	7.14***
DSM Anxiety	**.57*****	.06	6.15***
DSM Oppositional	.02	**.78*****	−10.73***
DSM Conduct Problems	.00	**.69*****	−8.87***
YOQ[e]			
Intrapersonal Distress	**.75*****	.35***	5.47***
Interpersonal Relations	.12	**.83*****	−9.61***
Social Problems	.02	**.69*****	−7.45***
Behavioral Dysfunction	.10	**.69*****	−6.73***
Somatic	**.48*****	.07	4.08***
SDQ[f]			
Emotional Symptoms	**.74*****	.06	8.06***
Conduct Problems	.03	.74***	−8.34***

Note: Correlations in bold are convergent validity coefficients, tests of association between scales of similar (internalizing vs. externalizing) content. Correlations not in bold are discriminant validity coefficients. *Z* values shown in column 4 test the significance of the difference between convergent and discriminant coefficients. BFS = Behavior and Feelings Survey; YSR = Youth Self-Report; *DSM* = *Diagnostic and Statistical Manual of Mental Disorders*; YOQ = Youth Outcome Questionnaire; SDQ = Strengths and Difficulties Questionnaire; CBCL = Child Behavior Checklist.

[a] *N* = 203.

[b] *N* = 148.

[c] *N* = 150.

[d] *N* = 222.

[e] *N* = 165.

[f] *N* = 167.

**p* < .05. ***p* < .01. ****p* < .001.

Study 3 Method

The Study 3 sample included 79 youths and their caregivers who had sought outpatient mental health treatment for an array of internalizing, externalizing, and other problems. Table 1 shows sample characteristics. Parents reported family income at $0–$39,999 for 65% of the sample, with 25% reporting $40,000–$79,999, 9% reporting $80,000–$119,999, and 1% reporting $120,000 or higher. Trained study staff administered the YSR and CBCL (described previously) and the 12-item BFS (described previously) to all 79 youths and caregivers separately.

Data Analytic Plan

Using baseline clinical data, we assessed the internal consistency and scale correlations of the 12-item BFS. We also investigated its convergent and discriminant validity with respect to the corresponding CBCL and YSR scales.

Study 3 Results and Discussion

Internal Consistency and Scale Correlations

The caregiver-report BFS showed good internal consistency for the total ($\alpha = .87$, M interitem $r = .33$), internalizing ($\alpha = .84$, M interitem $r = .48$), and externalizing ($\alpha = .94$, M interitem $r = .73$) scales. Internal consistency by youth report was similarly good: total ($\alpha = .87$, M interitem $r = .36$), internalizing ($\alpha = .91$, M interitem $r = .64$), and externalizing ($\alpha = .89$, M interitem $r = .58$). Regarding scale-scale correlations, by caregiver-report, internalizing and externalizing scales were uncorrelated with one another ($r = .08$, *ns*), but both were highly correlated with total BFS score (total-internalizing $r = 0.61$, $p < .001$; total-externalizing $r = .74$, $p < .001$). On the youth-report BFS, the internalizing and externalizing scales were weakly correlated ($r = .23$, $p = .039$), but both were highly correlated with the total BFS score (total-internalizing $r = .77$, $p < .001$; total-externalizing $r = .80$, $p < .001$).

Convergent and Discriminant Validity

Results of convergent/discriminant validity analyses involving the BFS and CBCL/YSR are shown in Table 4, with convergent coefficients in bold and discriminant coefficients not bolded. All z values were significant, indicating strong discriminant validity of BFS internalizing and externalizing scales. The same-informant associations between BFS total and CBCL/YSR total scores were high by both caregiver ($r = .61$, $p < .001$) and youth ($r = .72$, $p < .001$) report, supporting convergent validity.

To summarize, Study 3 examined the BFS when administered with no additional items, and findings showed convergent and discriminant validity in relation to the CBCL and YSR. These results essentially replicated the corresponding results from Study 2, but in a new clinical sample and with the final 12-item version of the BFS.

STUDY 4: PERFORMANCE OF THE BFS AS A PROGRESS MONITORING INSTRUMENT

In Study 4 we investigated the performance of the BFS in measuring progress over time, across repeated occasions during intervention. In particular, we examined (a) sensitivity of the BFS in detecting intraindividual change throughout treatment and (b) criterion validity in relation to an established progress monitoring measure, the Top Problems Assessment (Weisz et al., 2011)—which has been used to measure outcome trajectories and has shown strong sensitivity to change during treatment, in at least three published trials (Chorpita et al., 2017; Weisz et al., 2017, 2012).

TABLE 4
Study 3 Correlations of 12-Item BFS Subscales for Youth- and Caregiver-Report with YSR and CBCL Subscales

Youth Self-Report	*BFS Youth Internalizing*	*BFS Youth Externalizing*	*Discriminant Z Score*
Internalizing	**.74*****	.27*	4.15***
Externalizing	.27*	**.80*****	−5.07***
DSM Affective Problems	**.67*****	.28*	3.22***
DSM Anxiety	**.61*****	.03	4.19***
DSM Oppositional	.15	**.79*****	−5.67***
DSM Conduct Problems	.22*	**.74*****	−4.48***

Child Behavior Checklist	*BFS Caregiver Internalizing*	*BFS Caregiver Externalizing*	*Discriminant Z Score*
Internalizing	**.64*****	−.16	5.67***
Externalizing	−.08	**.81*****	−7.44***
DSM Affective Problems	**.54*****	.11	3.04**
DSM Anxiety	**.44*****	−.19	4.10***
DSM Oppositional	−.11	**.81*****	−7.63***
DSM Conduct Problems	−.18	**.75*****	−7.12***

Note: Correlations in bold are convergent validity coefficients, tests of association between scales of similar (internalizing vs. externalizing) content. Correlations not in bold are discriminant validity coefficients. *Z* values shown in column 4 test the significance of the difference between convergent and discriminant coefficients. BFS = Behavior and Feelings Survey; YSR = Youth Self-Report; CBCL = Child Behavior Checklist; *DSM* = *Diagnostic and Statistical Manual of Mental Disorders*.

*$p < .05$. **$p < .01$. ***$p < .001$.

Study 4 Methods

Participants were 95 youths receiving outpatient treatment from clinicians in school-based mental health services (18 schools in four urban school districts), plus the caregivers of these youths. Table 1 shows sample characteristics. Caregiver reports put family income for 36% at \$0–\$39,999, 22% at \$40,000–\$79,999, 16% at \$80,000–\$119,999; and 26% at \$120,000 or higher. Measures were administered at baseline, posttreatment, and weekly during treatment. Analyses used all cases with complete data, which we defined as four or more repeated observations collected from both parent and child. In practice, parents completed an average of 22.2 repeated assessments ($SD = 7.2$, range = 6–37) and youths completed 20.7 ($SD = 7.1$, range = 5–35).

Measures

The 12-item BFS (described previously) was administered to youths and caregivers weekly throughout treatment. In addition, the Top Problems Assessment (TPA; Weisz et al., 2011; used to monitor trajectories of change in multiple youth psychotherapy studies—e.g., Chorpita et al., 2017,; Weisz et al., 2012) was also administered weekly; it was used as an index of criterion validity. Youths and their caregivers identified their top problems at pretreatment and rated their severity each week throughout treatment (see Study 1 procedures). Mean TPA scores were calculated as the average of the three problem severity ratings for each informant and time point. Thus, TPA scores can be interpreted as a time-varying index of severity on the problems that matter most to the caregiver and youth, and therefore as an appropriate criterion for validity as a progress monitoring instrument. Evidence for test–retest reliability, convergent and discriminant validity, sensitivity to change, and slope-to-slope correlations with standardized measures all support the TPA's psychometric strength and utility for assessing trajectories of change during treatment (Weisz et al., 2011).

Data Analytic Plan

Mixed effects regression models were estimated to assess the performance of the BFS, relative to the TPA, in monitoring clinical change during treatment. Following previous youth intervention studies conducted with similar samples (Chorpita et al., 2013; Weisz et al., 2012), we used the natural logarithm of days since baseline as our metric of time. Trajectories are decomposed into a model for the means (fixed intercepts and log-linear slopes describing average trajectories over time) and a model for the variance (random effects and residual variance around those average trajectories). Fixed effects are interpreted such that higher intercept values represent greater severity at baseline, and more negative slope values indicate faster clinical improvement. Models were estimated in SAS Version 9.4 analyzing data in a double-stacked long data set format, using restricted ML estimation. Overall, this analytic plan allows for the estimation of parallel process models (i.e., trajectories of multiple outcome variables simultaneously), producing unbiased estimates while accommodating unbalanced and incomplete longitudinal data (Hoffman, 2015).

Analyses were conducted in two phases. First, we estimated the unconditional models for all TPA and BFS outcomes separately to describe the overall growth patterns on each variable. Second, we estimated parallel process models to examine relations between the TPA total trajectories and each of the BFS scale trajectories. These models yielded several cross-variable correlation terms, two of which are of primary interest here: (a) slope–slope correlations, or the average association between two scale trajectories, and (b) residual correlations, which indicate the average association between two scales' patterns of deviations from model-predicted values. Comparisons between TPA and BFS Total Problems were most relevant to the Study 4 questions, given the relative breadth of these two measures; the top problems varied widely in how much internalizing and externalizing were included. Nonetheless, we provide comparisons of TPA with BFS internalizing and BFS externalizing in order to present a complete picture.

Study 4 Results and Discussion

Unconditional Models

Results of the unconditional models are presented in Table 5. Across both informants on the BFS, trajectories started in the moderate to high range at baseline and showed decreases over time. These declining slopes were statistically significant for all scales except youth-report externalizing, which showed a marginal decline. (Youths had reported no clinical elevation in externalizing at baseline, so there was little room for improvement during treatment.) TPA trajectories followed a similar declining trajectory for both caregiver and youth report, with significant variability around the average estimates, as anticipated. There were also significant negative slope-intercept correlations on all measures by both informants, indicating that greater severity at baseline predicted faster trajectories of improvement over time, as is often found in treatment outcome research. Overall, these models show that BFS scales, like TPA scores, captured a pattern of intraindividual change—and interindividual variability around this average trajectory—reflecting clinical change over time.

Parallel Process Models

When the unconditional models just reported were combined in parallel process models, the results did not change appreciably from those reported in Table 5 and interpreted earlier; accordingly, we focus here only on the cross-variable correlation terms of interest. Caregiver-reported TPA slopes were strongly correlated with slopes for caregiver-reported BFS internalizing (r = .57, p < .001), externalizing (r = .54, p < .001), and total problems (r = .72, p < .001). Similarly, youth-reported TPA slopes were strongly associated with slopes for youth-reported BFS internalizing (r = .58, p < .001), externalizing (r = .52, p = .001), and total problems (r = .60, p < .001). In addition, the residual covariance terms between the BFS scales and TPA scale trajectories yielded significant correlations (all ps < .001) by both caregiver report (internalizing r = .41, externalizing r = .45, total r = .54) and youth report (internalizing r = .39, externalizing r = .36, total r = .45). From a conceptual and clinical perspective, these residual correlations suggest that on any occasion where a TPA score is higher or lower than predicted by the model (e.g., when there is a sudden gain or setback), the observed BFS scores also show a corresponding deviation from the model-predicted scores.

TABLE 5
Study 4 Results of Unconditional Univariate Mixed Models for BFS and TPA Trajectories Over the Course of Treatment

	Caregiver				*Youth*			
Models and Effects	*Est*	*SE*	p	r	*Est*	*SE*	p	r
BFS Internalizing								
Model for the Means								
Intercept	8.88	0.55	< .001		7.42	0.57	< .001	
Linear Slope (Log Days)	−0.70	0.11	< .001		−0.58	0.11	< .001	
Model for the Variance								
Intercept Variance	23.29	4.16	< .001		25.30	4.58	< .001	
Slope Variance	−2.43	0.70	.001		−2.54	0.74	.001	
Intercept-Slope Covariance	0.88	0.17	< .001	−.54	0.84	0.17	< .001	−.55
Residual Variance	8.98	0.29	< .001		8.96	0.30	< .001	
BFS Externalizing								
Model for the Means								
Intercept	10.49	0.69	< .001		6.06	0.60	< .001	
Linear Slope (Log Days)	−0.61	0.13	< .001		−0.23	0.13	.082	
Model for the Variance								
Intercept Variance	38.45	6.62	< .001		29.21	5.01	< .001	
Slope Variance	−3.91	1.04	< .001		−3.60	0.91	< .001	
Intercept-Slope Covariance	1.19	0.23	< .001	−.58	1.30	0.24	< .001	−.58
Residual Variance	11.23	0.36	< .001		8.20	0.28	< .001	
BFS Total								
Model for the Means								
Intercept	19.37	0.91	< .001		13.47	0.98	< .001	
Linear Slope (Log Days)	−1.31	0.19	< .001		−0.81	0.22	< .001	
Model for the Variance								
Intercept Variance	62.24	11.41	< .001		75.87	13.27	< .001	
Slope Variance	−4.50	1.88	.017		−8.35	2.38	.001	
Intercept-Slope Covariance	2.45	0.49	< .001	−.36	3.54	0.64	< .001	−.51
Residual Variance	26.31	0.85	< .001		23.59	0.79	< .001	
TPA Total								
Model for the Means								
Intercept	3.26	0.08	< .001		2.75	0.09	< .001	
Linear Slope (Log Days)	−0.31	0.02	< .001		−0.32	0.02	< .001	
Model for the Variance								
Intercept Variance	0.38	0.09	< .001		0.49	0.11	< .001	
Slope Variance	−0.05	0.02	.019		−0.04	0.02	.051	
Intercept-Slope Covariance	0.04	0.01	< .001	−.40	0.03	0.01	< .001	−.36
Residual Variance	0.37	0.01	< .001		0.43	0.01	< .001	

Note: Parameter estimates describe the log-linear trajectories of change for each progress monitoring scale by each informant (eight models estimated). Each model is decomposed into two parts: (a) the *model for the means*, which estimates the average score at baseline (fixed intercept) and average rate of change over time (log-linear slope), and (b) the *model for the variance*, which characterizes patterns of variation, covariation, and residual variance by which individuals deviate from these average trajectories. For interpretability, covariance terms are also reported as correlations. BFS = Behavior and Feelings Survey; TPA = Top Problems Assessment.

To summarize, all BFS scale scores showed moderate to strong associations with the TPA scores over time—in terms of both overall trajectory of change and session-to-session deviations from that trajectory—thus supporting the criterion validity and clinical utility of the BFS as a progress monitoring instrument.

GENERAL DISCUSSION

We developed a brief measure for use in frequent monitoring of treatment response by young people throughout episodes of therapy. Such a measure could be used during treatment as a part of MBC (Scott & Lewis, 2017), to inform clinical judgments about how well interventions are working and whether midcourse adjustments may be needed, and after treatment to plot trajectories of change for individuals and groups. Our findings indicate that the BFS has the characteristics needed for these applications. Across a series of studies, the BFS showed robust factor structure, internal consistency, test–retest reliability, convergent and discriminant validity in relation to three well-established symptom measures, and slopes of change indicating viability as a progress monitoring instrument when used frequently throughout full episodes of treatment.

The BFS was designed to be both practice friendly and empirically sound. Its brevity—less than 1 min is required to rate the 12 items—should help minimize measurement burden, increasing the likelihood that youths and caregivers will complete it regularly throughout treatment. The focus on internalizing and externalizing, the two most thoroughly documented dimensions of youth psychopathology, locates the BFS within a strong empirical tradition and encompasses problems that are highly prevalent in clinical care settings. Making the BFS free to all should enhance prospects for routine use by clinicians, service programs, and researchers in times of limited funding. The combination of brevity, accessibility, and psychometric support could make the BFS a useful tool for everyday clinical practice and research. In both contexts, the focus of the BFS on internalizing and externalizing may omit highly specific problems of great personal relevance to youths and to caregivers. As noted previously, these may be identified and measured repeatedly through such idiographic means as the "top problems assessment" (Weisz et al., 2011); an especially sensitive and clinically valuable form of MBC may involve combining such highly personalized idiographic measurement with the kind of nomothetic measurement represented by the standardized BFS, which provides a common metric for parents and youths, and across multiple clients, using the same set of items.

Our examination of the BFS items and scores within three clinical samples (i.e., Studies 2–4) found considerable similarity across the samples (see Supplemental Materials). Consistent with the literature (e.g., De Los Reyes et al., 2015), caregivers reported higher levels of externalizing than youths, and cross-informant correlations were medium to large for externalizing (*r*s = .33–.52), null to medium for internalizing (*r*s = .00–.39), and small to medium for total problems (*r*s = .15–.36). Further, all 12 items were sensitive to elevated symptoms, with item-level means and standard deviations generally between 1 and 2, and spanning the entire scale range from 0 to 4, with no evidence of a floor effect in any sample. In all three samples, BFS scale scores showed few and mixed associations with age, gender, and ethnicity; given the differences in size and makeup of the three samples, these findings are best seen as preliminary, warranting more definitive assessment in future research with samples selected for that purpose.

The potential strengths of the BFS, and of the research in clinical contexts used to examine its psychometrics, should be considered in light of certain limitations of the measure and the research. First, the BFS is not intended to be an all-purpose or comprehensive measure. Measure brevity is achieved through focus, and our focus on internalizing and externalizing problems necessarily omitted other problem domains that may warrant clinical attention. MBC with treatment that uniformly targets other problem domains (e.g., autism spectrum, posttraumatic stress disorder) may be guided by measures that focus precisely on those domains. A limitation of the BFS, as a new measure, is that it lacks the extensive base of accumulated evidence that has created a rich tapestry of norms and applications for the CBCL/YSR, YOQ, SDQ, and other venerable measures of youth psychopathology. Such developments require many years of data collection and analysis beyond initial measure validation. Another limitation is that our studies focused only on caregivers who have mental health concerns about their children (in Study 1) and caregivers and youths in outpatient care (in Studies 2, 3, and 4). More will need to be learned in the future about functioning of the BFS with reporters other than youths and caregivers and in other service settings (e.g., inpatient and primary care). Finally, because the BFS is composed of a fixed set of standard items, it cannot provide for individualized assessment of person-specific functional problems that may be especially important to each youth and caregiver. As noted previously, this limitation could be addressed by combining the BFS with the TPA (Weisz et al., 2011), described earlier, which entails weekly ratings by each youth and caregiver on the three "top problems" each identified at treatment outset. Weekly TPA ratings can add personalized functional assessment to MBC, complementing the standardized assessment of the BFS (and adding less than 15 s to the assessment, on average). Such a combination of standardized and personalized MBC has been used in previous research (e.g., Chorpita et al., 2017, 2017; Weisz et al., 2012), both to guide clinicians during treatment and to generate data on trajectories of change, for use in outcome analyses.

Future research might also examine whether different ways of deploying MBC—with BFS, or BFS combined with TPA—might differ in their impact on psychotherapy outcome. Evidence reviewed in the introduction (e.g., Fortney et al., 2017) indicates that routine MBC feedback to treating clinicians has been associated with improved treatment effects. A useful question for the future is whether outcomes might be improved further if caregivers, and perhaps even youths, were to not only complete weekly ratings but also see their own ratings—plotted across the weeks, to help them monitor their own progress. In this and other ways, the BFS might be used to test new ideas for involving youths and families and expanding their treatment benefit.

SUPPLEMENTARY MATERIAL

Supplemental data for this article can be accessed on the publisher's website.

ACKNOWLEDGMENTS AND FUNDING

We are grateful to the youths, caregivers, and clinical staff and administrators who participated in these studies and to the Annie E. Casey Foundation (grant 211.0004) and the Norlien Foundation for their support. The Brief Problem Checklist (BFS), copyrighted by Harvard University, is available free of charge to all who wish to use it. The Youth- and Caregiver-report forms of the BFS can be found at this site: http://weiszlab.fas.harvard.edu/

REFERENCES

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs: General and Applied*, *80*(7), 1–37, (Whole No. 615). ISSN: 0096-9753; PMID: 5968338 Version:1, . doi:10.1037/h0093906

Achenbach, T. M., Conners, C. K., Quay, H. C., Verhulst, F. C., & Howell, C. T. (1989). Replication of empirically derived syndromes as a basis for taxonomy of child/adolescent psychopathology. *Journal of Abnormal Child Psychology*, *17*(3), 299–323. doi:10.1007/BF00917401

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Youths, Youth and Families.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders*, 4th edition. Washington, DC: American Psychiatric Association.

Bickman, L., Douglas Kelley, S., Breda, C., de Andrade, A. R., & Riemer, M. (2011). Effects of routine feedback to clinicians on mental health outcomes of youths: Results of a randomized trial. *Psychiatric Services*, *62*(12), 1423–1429. doi:10.1176/appi.ps.002052011

Burlingame, G. M., Mosier, J. I., Wells, M. G., Atkin, Q. G., Lambert, M. J., Whoolery, M., & Latowsky, M. (2001). Tracking the influence of mental health treatment: The development of the youth outcome questionnaire. *Clinical Psychology and Psychotherapy*, *8*(5), 315–334. doi:10.1002/cpp.315

Chorpita, B. F., Daleiden, E. L., Park, A. L., Ward, A. M., Levy, M. C., Cromley, T., … Krull, J. L. (2017). Child STEPs in California: A cluster randomized effectiveness trial comparing modular treatment with community implemented treatment for youth with anxiety, depression, conduct problems, or traumatic stress. *Journal of Consulting and Clinical Psychology*, *85*(1), 13–25. doi:10.1037/ccp0000133

Chorpita, B. F., Reise, S., Weisz, J. R., Grubbs, K., Becker, K. D., & Krull, J. L.; Research Network on Youth Mental Health. (2010). Evaluation of the Brief Problem Checklist: Child and caregiver interviews to measure clinical progress. *Journal of Consulting and Clinical Psychology*, *78*(4), 526–536.

Chorpita, B. F., Weisz, J. R., Daleiden, E. L., Schoenwald, S. K., Palinkas, L. A., Miranda, J., Higa-McMillan, C. K., Nakamura, B. J., Austin, A. A., Borntrager, C., Ward, A. M., Wells, K. C., & Gibbons, R. D., & the Research Network on Youth Mental Health. (2013). Long term outcomes for the Child STEPs randomized effectiveness trial: A comparison of modular and standard treatment designs with usual care. *Journal of Consulting and Clinical Psychology*, *81*(6),999–1009. doi:10.1037/a0034200.

Chorpita, B. F., & Weisz, J. R. (2009). *Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems (MATCH-ADTC)*. Satellite Beach, FL: PracticeWise, LLC.

Cicchetti, D., & Toth, S. L. (1991). *Internalizing and externalizing expressions of dysfunction*. Rochester Symposium on Developmental Psychopathology, Vol. 2. Rochester, NY: University of Rochester Press.

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory*. New York, NY: Guilford Press.

De Los Reyes, A., Augenstein, T. M., & Aldao, A. (2017). Assessment issues in child and adolescent psychotherapy. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 537–554). New York, NY: Guilford Press.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, G. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858–900. doi:10.1037/a0038498

Dorsey, S., Berliner, L., Lyon, A. R., Pullmann, M. D., & Murray, L. K. (2014). A statewide common elements initiative for children's mental health. *Journal of Behavioral Health Services and Research*, *43*(2), 46–26. doi:10.1007/s11414-014-9430-y

Dunn, T. W., Burlingame, G. M., Walbridge, M., Smith, J., & Crum, M. J. (2005). Outcome assessment for children and adolescents: Psychometric validation of the Youth Outcome Questionnaire 30.1. *Clinical Psychology & Psychotherapy*, *12*(5), 388–401. doi:10.1002/cpp.461

Fortney, J. C., Unutzer, J., Wren, G., Pyne, J. M., Smith, G. R., Schoenbaum, M., & Harbin, H. T. (2017). A tipping point for measurement-based care. *Psychiatric Services*, *68*(2), 179–188. doi:10.1176/appi.ps.201500439

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(11), 1337–1345. doi:10.1097/00004583-200111000-00015

Gross, T. J., Hurley, K. D., Lambert, M. C., Epstein, M. H., & Stevens, A. L. (2015). Psychometric evaluation of the Symptoms and Functioning Severity Scale (SFSS) short forms with out-of-home care youth. *Child Youth Care Forum*, *44*, 239–249. doi:10.1007/s10566-014-9280-z

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151. doi:10.1177/001316446002000116

Little, R. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: the eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–20.

Mash, E. J., & Wolfe, D. A. (2013). *Abnormal child psychology*. Boston, MA: Cengage Learning.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175. doi:10.1037/0033-2909.111.1.172

Ng, M. Y., & Weisz, J. R. (2016). Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry*, *57*(3), 216–236. doi:10.1111/jcpp.12470

Quay, H. C. (1979). Classification. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (2nd ed., pp. 1–42). New York, NY: Wiley.

R Core Team. (2015). *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. R package version 1.8.10. Northwestern University, Evanston, Illinois, USA. https://CRAN.R-project.org/package=psych

Revelle, W., & Rocklin, T. (1979). Very Simple Structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403–414. doi:10.1207/s15327906mbr1404_2

Scott, K., & Lewis, C. (2017). Using measurement-based care to enhance any treatment. *Cognitive & Behavioral Practice*, *22*(1), 49–59. doi:10.1016/j.cbpra.2014.01.010

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, *78* (3), 298–311. doi:10.1037/a0019247

Staller, J. A. (2006). Diagnostic profiles in outpatient child psychiatry. *American Journal of Orthopsychiatry*, *76*(1), 98–102. doi:10.1037/0002-9432.76.1.98

Vostanis, P. (2006). Strengths and Difficulties Questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, *19*(4), 367–372. doi:10.97/01.yco.0000228755.72366.05

Weisz, J. R., Bearman, S. K., Santucci, L., & Jensen-Doss, A. (2017). Initial test of a principle-guided approach to transdiagnostic psychotherapy with children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *46*(1), 44–58. doi:10.1080/15374416.2016.1163708

Weisz, J. R., Chorpita, B. F., Frye, A., Ng, M. Y., Lau, N., & Bearman, S. K.; Research Network on Youth Mental Health. (2011). Youth top problems: Using idiographic, consumer-guided assessment to identify treatment needs and track change during psychotherapy. *Journal of Consulting and Clinical Psychology*, *79*(3), 369–380.

Weisz, J. R., Chorpita, B. F., Palinkas, L. A., Schoenwald, S. K., Miranda, J., & Bearman, S. K.; Research Network on Youth Mental Health. (2012). Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. *Archives of General Psychiatry*, *69*(3), 274–282.

Wolpert, M., Cheng, H., & Deighton, J. (2015). Measurement issues: Review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO-their strengths and limitations for clinical use and service evaluation. *Child and Adolescent Mental Health*, *20*(1), 63–70. doi:10.1111/camh.12065

Youngstrom, E., Findling, R., & Calabrese, J. (2003). Who are the comorbid adolescents? Agreement between psychiatric diagnosis, youth, parent, and teacher report. *Journal of Abnormal Child Psychology*, *31*(3), 231–245. doi:10.1023/A:1023244512119