Youth top problems: Using idiographic, consumer-guided assessment to identify treatment needs and to track change during psychotherapy.

In Press, *Journal of Consulting and Clinical Psychology*

Youth Top Problems: Using Idiographic, Consumer-Guided Assessment to Identify Treatment Needs and Track Change during Psychotherapy

John R. Weisz

Harvard University and Judge Baker Children's Center

Bruce F. Chorpita

University of California at Los Angeles

Alice Frye

Wellesley Centers for Women

Mei Yi Ng and Nancy Lau

Harvard University

Sarah Kate Bearman, Ana Ugueto, and David A. Langer

Judge Baker Children's Center and Harvard University

Kimberly E. Hoagwood

Columbia University

The Research Network on Youth Mental Health

Running head: Assessing Youth Top Problems

Abstract

Objective. To complement standardized measurement of symptoms, we developed and tested an efficient strategy for identifying (before treatment) and repeatedly assessing (during treatment) the problems identified as most important by caregivers and youths in psychotherapy. ***Method***. 178 outpatient-referred youths, aged 7-13, and their caregivers separately identified the three problems of greatest concern to them at pre-treatment, and then rated the severity of those problems weekly during treatment. The Top Problems measure thus formed was evaluated for (a) whether it added to the information obtained through empirically-derived standardized measures (e.g., the Child Behavior Checklist [CBCL] and Youth Self-Report [YSR]), and (b) whether it met conventional psychometric standards. ***Results***. The problems identified were significant and clinically-relevant; most matched CBCL/YSR items while adding specificity. The top problems also complemented the information yield of the CBCL/YSR; for example, for 41% of caregivers and 79% of youths the identified top problems did not correspond to any items of any narrowband scales in the clinical range. Evidence on test-retest reliability, convergent and discriminant validity, sensitivity to change, slope reliability, and the association of Top Problems slopes with standardized measure slopes supported the psychometric strength of the measure. ***Conclusions***. The Top Problems measure appears to be a psychometrically sound, client-guided approach that complements empirically-derived standardized assessment; the approach can help focus attention and treatment planning on the problems youths and caregivers consider most important and generate evidence on trajectories of change in those problems during treatment. [239 words]

Keywords: Assessment, Top Problems, Youth (Children, Adolescents), Psychotherapy

Youth Top Problems: Using Idiographic, Consumer-Guided Assessment to Identify Treatment Needs and Track Change during Psychotherapy

It is common practice at the beginning of everyday youth psychotherapy for clinicians to ask clients—i.e., youths and their caregivers—what problems they are most concerned about and would like to address in treatment. The resulting discussion can help build rapport, initiate a working alliance, and identify therapy goals. However, the discussion of problems may not be structured consistently across cases, the information about client-identified problems may not be used in a very systematic way thereafter, and progress regarding these initial concerns may not actually be assessed. This paper focuses on whether the widespread practice of having clients identify top problems can be structured to form a psychometrically sound, client-guided approach to evidence-based assessment, one that might complement more standardized methods.

Such an approach could have value for clinical science and practice, potentially enriching the interplay of evidence-based assessment and evidence-based treatment that so many have urged (see Achenbach, 2005; Hunsley & Mash, 2007; Weisz, Chu, & Polo, 2004). The ideal may be a kind of "assessment-intervention dialectic" (Weisz et al., 2004) in which assessment is used to plan treatment, modify it in response to changes in client functioning, and determine when treatment should end. To date, most research on the assessment part of this dialectic has focused on empirically-derived standardized measures in which clients answer a fixed series of questions and responses are scored according to a fixed set of dimensions or scales (e.g., Child Behavior Checklist [CBCL] and Youth Self-Report [YSR], Achenbach & Rescorla, 2001). This valuable approach has added greatly to the rigor of clinical assessment. Perhaps assessment could be further enriched by a complementary focus on identifying and monitoring change in the problems clients themselves identify as important (cf. Cone, 1999; Kazdin, 2000).

Assessment of client-identified problems could support clinical practice in several ways: (a) adding specificity to problems that are identified generically in standardized measures; (b) focusing therapist attention on client concerns that would not be identified via standard use of standardized measures; (c) identifying specific client priorities within a large array of problems, whether evident in standardized measures or not; (d) giving clients a voice in shaping the agenda and goals of treatment; (e) enhancing rapport and alliance between clients and clinicians; (f) providing foci for ongoing assessment during treatment, and thus a way to gauge whether treatment is impacting the problems clients consider most important; (g) informing decisions about when to end treatment, based in part on whether client concerns have been successfully addressed; and (h) using an approach that can fit into everyday practice because it builds on an already-widely-used procedure—i.e., identifying client concerns at the beginning of treatment. Using top problem assessment to pursue these goals would be consistent with calls for more idiographic approaches (e.g., Barlow & Nock, 2009) and client-guided methods of clinical assessment (e.g., Eifert, Evans, & McKendrick, 1990; Hoagwood et al., 2010)

Client-guided assessment may be particularly important in youth treatment because it typically entails clinical work with two clients—youth and caregiver. For the clinician seeking to engage both parties, it may be critical to know what problems each sees as most important, and to assess treatment progress in relation to those problems. Caregiver perspective is key because it is the caregivers who typically initiate the treatment, and they often know of issues (e.g., conduct problems at home) that may be less evident or less distressing to youths. Moreover, a failure to address their paramount concerns may undermine the caregiver support needed for treatment success (e.g., transporting the youth to the clinic, encouraging the youth to participate and cooperate, keeping the clinician informed about events at home). Indeed, Bannon and McKay

(2005) found that families whose clinical care did not match what caregivers had requested for their youths at intake ended treatment earlier than families for whom the services did match caregiver wishes. Similarly, in marital and family therapy, the "fit" of treatment to specific client needs and expectations has been found to account for 35% of the variance in client outcome (Crane, Griffin, & Hill, 1986).

The perspective of the youth in treatment is also valuable. Youths may identify problems (e.g., specific fears, depressive symptoms) of which caregivers are unaware. Moreover, youth engagement and attentiveness during sessions, participation in session activities, frank discussion with the therapist, and homework completion may be jeopardized if treatment does not address problems the youths consider most important. Indeed, because boys and girls rarely self-refer, identifying concerns that youngsters themselves are motivated to work on may be critical to success. The importance of obtaining the perspective of both caregiver *and* youth is underscored by the common finding that these two perspectives are poorly correlated (see Meyer et al., 2001); thus, the perspective of one cannot be assumed to represent the other's point of view.

For all these reasons, having youths and caregivers identify the problems most important to them can be useful at the outset of treatment. The utility of this information may be greatly magnified if severity ratings on the identified problems can be obtained frequently during treatment. Such assessment can help meet the expanding need for brief measures of treatment response that can be used to guide treatment planning and case supervision, and to measure outcome trajectories (cf. Chamberlain & Reid, 1987; Webster-Stratton & Spitzer, 1991).

Such frequent measurement can serve treatment research in several ways. First, it is ideal for the new generation of methods for modeling change during treatment (e.g., Raudenbush & Bryk, 2002). Second, improvement occurring early in treatment (e.g., Ilardi & Craighead, 1994) may

go undetected by measurement approaches delivered at post-treatment only, which may in turn increase Type II error in tests of group differences (see e.g., Weisz et al., 2009). Third, increased use of effectiveness designs that pit time-limited evidence-based treatments against usual care of uncontrolled length (see Weisz, Jensen-Doss, & Hawley, 2006), with group differences in treatment dose and duration highly likely, make *slope* across multiple assessment points a more sensitive index of outcome than post-treatment measurement alone. Fourth, frequent assessment during treatment can provide raw material for mediation tests, and thus contribute to identifying mechanisms of change (see e.g., Kazdin, 2007; Weersing & Weisz, 2002).

Frequent assessment can also be valuable in clinical practice. A therapist's capacity to plan treatment and adjust intervention procedures appropriately during care can be enhanced by ongoing information about client response (e.g., Chorpita, Bernstein, & Daleiden, 2008; Lambert, Harmon, Slade, Whipple, & Hawkins, 2005; Weisz et al., 2004; Weisz & Chorpita, in press). Moreover, clients often unilaterally end treatment without notice and thus are unavailable for termination debriefing or post-treatment assessment. In such cases, if frequent measurement is in place throughout treatment, the last measurement automatically becomes the end-of-treatment assessment, and the full trajectory-of-change is automatically documented across the prior measurement points.

The need for brief, frequent assessment with standard items lists has been addressed via measures focused on child conduct (e.g., Chamberlain & Reid, 1987), discipline at home (Webster-Stratton & Spitzer, 1991), and youth internalizing and externalizing problems (Chorpita, Reise, Weisz, Grubbs, Becker, & Krull, 2010). Measures like these, with uniform lists of items focused on a fixed set of dimensions, might be usefully complemented by an idiographic

approach that adds client-identified top problem ratings to the measurement model. However, questions arise regarding such an idiographic approach.

Question 1: Hasn't this approach been tried before? This exact approach has not, but at least three kinds of prior work are relevant. First, several researchers have obtained reasons given by youths and caregivers for seeking treatment, and have classified these reasons for degree of fit to items on standardized measures (e.g., Hawley & Weisz, 2003; Lambert, Rowan, Lyubansky, & Russ, 2002; Yeh & Weisz, 2001; Weisz & Weiss, 1991); unlike these efforts, our approach involves (a) obtaining severity ratings on the problems when identified, and (b) continuing to obtain the ratings during treatment, to monitor change. Second, a family of methods exemplified by *goal attainment scaling* (GAS; Kiresuk & Sherman, 1968) has been used to track progress toward treatment goals. In GAS, client goals are identified by individuals, clinicians, or a committee, and standardized scores are used to measure progress toward the goals. GAS has been criticized for several problems, none relevant to our proposed approach (see Cytrynbaum, Ginath, Birdwell, & Brandt, 1979; Lambert et al., 1986; MacKay, Somerville, & Lundie, 1996). Unlike GAS, our approach (a) does *not* arbitrarily weight problems or standardize scores, (b) was *not* designed to be a standardized measure, (c) does *not* use raters who are themselves the treatment providers, and (d) *does* show a significant association with other more standard measures of improvement (see Results, below). However, the top problems approach does share several of the strengths of GAS summarized by MacKay et al. (1996)—e.g., it supports systematic evaluation and monitoring of change, involves clients in the process, and focuses service efforts on consumers' goals. Finally, the top problems approach proposed here does operate in the way Lambert et al. (1986) proposed the GAS should be used—i.e., "in conjunction with standard scales applied to all patients" (p. 192). In a third line of research, closest

conceptually to what is proposed here, Doss, Simpson, and Christensen (2004) assessed partners' reasons for seeking marital therapy, and Doss, Thum, Sevier, Atkins, and Christensen (2005) obtained husbands' and wives' ratings on these reasons at four assessments spaced about 13 weeks apart. These ratings showed excellent sensitivity to change during treatment, and they figured importantly in outcome analyses. Our approach is similar in concept but it differs in that (a) it focuses on youths in treatment, (b) it involves a denser assessment schedule, and (c) it includes extensive assessment of psychometric properties, including association of top problem ratings with a widely-used standardized measure and theoretically significant dimensions of psychopathology (i.e., internalizing and externalizing--see below).

Question 2: Might the youth and caregiver top problems be non-clinical in nature, or inappropriate targets for treatment? This question was addressed in three ways. First, to set an appropriately clinical context for the questions, youths and caregivers were asked to identify their top problems after they had completed a diagnostic assessment. Second, youths and caregivers ranked the problems in terms of their priorities for treatment, and the three that were ranked most important were used. Third, to investigate the question through data analysis, a coding system was used to assess the degree to which top problems matched the clinical problem items of the CBCL and YSR.

Question 3: Don't new problems arise during treatment? This seems likely to occur, just as new DSM-IV diagnoses and new clinical range scores on standardized measures can arise during treatment (see e.g. Goodyer, Herbert, Secher, & Pearson, 1997; Kovacs, Obrosky, & Sherrill, 2003; Ollendick & King, 1994). In fact, if treatment is successful in addressing the referral concerns, the original problems (and diagnoses and clinical scale scores) *should* diminish in severity and importance relative to new concerns that have not been addressed in treatment. In

principle, one might change treatment goals each time new problems are identified, but this could lead to rather chaotic treatment. Instead, a case can be made for identifying the primary concerns that lead to treatment, adopting these as treatment targets, and monitoring change over time to track whether the targets are being addressed successfully. This is the procedure typically followed in well-designed RCTs, and arguably in the best evidence-informed clinical care. In many practice settings a required first step is creating a kind of 'service contract' by having clients identify their main problems. The top problems assessment approach is essentially a way to monitor performance on the contract.

Answers to the three preceding questions notwithstanding, the potential of top problems assessment for research and clinical applications depends on (a) whether the information is clinically relevant and adds usefully to standardized assessment and (b) how top problem ratings perform psychometrically. Assessing the psychometrics of top problem ratings poses a challenge: the lack of a uniform list of items—the typical focus in psychometric assessment of standardized measures—means that the criterion measures with which these ratings would be expected to correlate differ across respondents. Nevertheless, if data analyses are structured properly, test-retest reliability, convergent validity, and discriminant validity may be quite relevant to idiographic problem lists (see e.g., Cone, 1999; Haynes & O'Brien, 2000). In addition, it is possible to examine criterion validity, focusing on the sensitivity of top problem ratings to change during treatment, in relation to change in well-established measures.

Accordingly, this study included assessment of (a) the clinical relevance and information value of top problem identification, and (b) the psychometrics of top problem severity ratings, in a sample of 178 youths in community-based outpatient treatment. In each case youth and caregiver were asked, separately at pre-treatment, to identify the top three problems for which

help was needed. The problems were then rated weekly during treatment. Clinical relevance of the identified problems was assessed by examining their degree of match to the problem items of widely-used standardized clinical measures. The question of whether identifying top problems complemented a standardized approach was answered by assessing whether the top problems noted would have been prioritized through standard scoring of standardized measures. And psychometrics of the Top Problems assessment was examined by focusing on (a) test-retest reliability at three points during treatment; (b) convergent validity—i.e., whether ratings on empirically meaningful clusters of top problems were positively associated with theoretically related dimensions of standardized measures; (c) discriminant validity-i.e., whether ratings on key clusters of top problems were *not* positively associated with theoretically *un*related standardized measures; and (d) criterion validity over time during treatment, as evidenced by sensitivity to change, slope reliability, and the correlation of top problem trajectories with standardized measure trajectories.

Method

Participants were 178 youths and their caregivers seeking outpatient treatment for an array of internalizing and externalizing problems. Participants were drawn from nine community outpatient programs providing treatment to children in office-based and school-based settings in two metropolitan areas. Treatment was provided by 85 therapists (76% master's degrees, 21% doctoral degrees, 2% bachelor's degree). Therapist-reported orientations were 35% cognitive behavioral, 21% eclectic, 12% psychodynamic, 8% family systems, 7% behavioral, and 17% other (e.g., Adlerian, Play Therapy). Mean treatment duration was 226 days (*SD* =145), 17 sessions (*SD* = 12).

For inclusion, youths were required to show at least borderline elevation (T>64) on at least one Internalizing or Externalizing narrowband scale of the CBCL or YSR (see below), and to either meet diagnostic criteria for a DSM-IV anxiety disorder, depressive disorder, or disruptive behavior disorder, or have caregiver- or youth-reported disturbances of anxiety, depression, or disruptive behavior that did not meet full diagnostic criteria. Youths with recent psychiatric hospitalizations or suicide attempts or with evidence of psychosis or pervasive developmental disorders were excluded. The criteria provided for a very broad array of youths (see below) but excluded those who might be unable to understand and accurately respond to self-report assessments. The study was conducted in compliance with an authorized Institutional Review Board. Trained project staff presented consent/ assent forms in writing and orally to caregivers and youths, respectively. Of those invited to participate, 62.46% agreed and signed consent/assent. Among those who agreed and signed, the attrition rate (failure to complete the phone assessments) was 8.43%.

Youths' ages ranged from 7.15 to 13.97 years, with a mean of 10.62 (*SD* = 1.81); 68% were boys; 44% were European-American, 32% multi-ethnic, 10% African-American, 7% Hispanic-American, 4% Asian-American, 2% Pacific Islander, and 2% other. Diagnoses based on structured interviews (see method below) spanned multiple internalizing and externalizing disorders, as is common in youth outpatient settings. The most common diagnostic categories were ADHD (55.62% of the sample), anxiety disorders (50.00%), oppositional defiant disorder (47.19%), major depressive disorder (24.16%), and dysthymic disorder (11.80%). These and the other disorders in the sample sum to more than 100% because comorbidity was substantial; 80.90% of the sample had two or more diagnoses, and the mean number of disorders per youth was 2.68 (SD=1.52).

Caregiver participants were mothers (biological, adoptive, or step-mothers; *n* = 152; 85.4%), grandparents (*n* = 14; 7.9%), fathers (*n* = 10; 5.6%), uncle (*n* = 1; 0.6%), and great-great aunt (*n* = 1; 0.6%). Some 42% were married, 23% divorced, 17% single parent, 8% separated, 6% living with partner, and 5% widowed. Modal education level was a high school diploma or equivalent. With household income grouped into $20K intervals, median income was in the $20,000-$39,000 range. Households averaged 3.81 family members (SD = 1.45).

Measures

The Children's Interview for Psychiatric Syndromes-Child and Parent Forms (ChIPS and ChIPS-P; Weller et al, 2000) are structured diagnostic interviews based on DSM-IV criteria and designed for ages 6-18 and caregivers. Symptoms are assessed via a yes/no format, with simple wording to enhance comprehension. Psychometric analyses have shown high test-retest reliability, moderate to high correlations with discharge diagnoses, and good agreement with other standardized diagnostic interviews. Five psychometric studies combined (Weller et al, 2000) showed overall sensitivity vis-à-vis clinician diagnoses at .66 for CHIPS and .83 for P-ChIPS, and overall specificity at .88 for ChIPS and .78 for P-ChIPS.

Youth Self-Report and Child Behavior Checklist (YSR and CBCL; Achenbach & Rescorla, 2001). The YSR and CBCL are parallel 118-item self-report and caregiver-report measures of youth behavioral and emotional problems. Youths and their caregivers rate each item 0 (Not True), 1 (Somewhat or Sometimes True), 2 (Very True or Often True). Both measures generate a total problems scale, broadband Internalizing and Externalizing syndrome scales, eight narrowband syndrome scales (e.g., Anxious-Depressed, Aggressive Behavior), and six DSM-oriented scales corresponding to diagnostic clusters; DSM scales relevant to this study were Affective Problems, Anxiety Problems, Oppositional Defiant Problems, and Conduct Problems.

The YSR also generates a Positive Qualities scale based on its 14 non-problem items (e.g., I like animals; see Rescorla et al., 2007). The CBCL and YSR are supported by extensive evidence encompassing reliability, validity, and clinical utility (see Achenbach & Rescorla, 2001).[1]

Brief Problem Checklist (BPC; Chorpita et al., 2010). The 12-item BPC was derived from the application of factor analysis and item response theory to YSR and CBCL data from 2332 youths and caregivers, to produce a brief measure that can be administered frequently to assess youth problems during treatment. The 12 items use the same 0-1-2 response format as the YSR and CBCL; they generate a Total Problems score plus factor analytically-derived Internalizing and Externalizing scores based on six items each. The Total, Internalizing, and Externalizing scores have shown strong reliability, internal consistency, and convergent and discriminant validity in relation to the corresponding and distinct YSR and CBCL scores and DSM-IV diagnoses. In longitudinal analyses the BPC significantly predicts change on other measures of youth symptoms and dysfunction, and estimates from random coefficient growth models have shown generally higher reliability estimates for slopes using weekly BPC scores than for slopes using scores from the full CBCL and YSR administered every 3 months (Chorpita et al., 2010). The findings show the BPC to be a psychometrically sound measure of core constructs assessed by the lengthier YSR and CBCL, with brevity that supports frequent administration.

Brief Symptom Inventory (BSI). Caregivers completed the BSI (Derogatis & Melisaratos, 1983), a 53-item self-report measure of caregiver general psychopathology and psychological distress that has shown good test-retest reliability, internal consistency, and convergent validity with more extensive established measures (see e.g., Derogatis & Melisaratos, 1983; Hafkenscheid, 1991).

Top Problems (TP) measure. The TP assessment was administered separately to youths and caregivers. To establish a serious clinical context and thus reduce the risk of trivial responses, TP assessment always followed the diagnostic assessment (see above). After the diagnostic questions, youths and caregivers were asked to list the problems they were most concerned about, the interviewer wrote these down in respondents' own words (e.g., "My mom and I argue a lot."), then asked whether there were other problems yet identified that should also go on the list. When the list was complete, the interviewer obtained severity ratings for each problem ["How big of a problem is this for you" (youth) or "…for her/him" (caregiver)?] on a scale of 0 ["not at all"] to 10 ["very, very much"]. Each youth and caregiver was given a list of all the problems s/he had identified and asked which one "is the biggest problem right now? Which of these is giving you [or youth's name] the most trouble right now? Which one is the most important to work on?" The problem thus identified was assigned rank #1; then the interviewer asked for the next biggest problem, and the next. This resulted in a ranked list of the top three problems identified by youth and by caregiver, which formed the TP measure.

Study Procedure and Weekly Assessments

When youths in the study age range were referred for treatment, they and their caregivers were told about the study. Those who were interested were screened and interviewed by project assessors, with study measures administered before treatment began. The BPC and TP measures were administered to youths and caregivers separately by phone, with a target interval of 7 days (exact intervals noted below). In each TP assessment, youth and caregiver (separately) were read their three top problems and asked to rate the severity of each on the 0-10 scale (see above). Assessors knew the name and gender of youth and parent but were blind to all clinical and study-related information about participants. Review of youth and caregiver ratings at the midpoint of

the study, call 13, showed that both groups used the full range (i.e., all 11 scale points) of the 0-10 scale (youth mean 3.37, SD 2.82; caregiver mean 5.16, SD 2.81). Across the 26 assessment points of the study, averaging across the three top problems, the percentage of youths using each of the 11 scale points ranged from 3.56% to 29.50; the percentages for caregivers ranged from 3.41% to 12.75%. For both youths and caregivers, the full raw distribution of ratings was used in all analyses at all time points.

Top Problems Coding, and Composite Internalizing, Externalizing and Total Scores

To facilitate psychometric assessment of the TP measure in relation to a well-established standardized measure, the top problems identified by caregivers and youths were coded using a system based on correspondence to CBCL/YSR items. The system was developed and reported by Weisz and Weiss (1991), Yeh and Weisz (2001), and Hawley and Weisz (2003), all of whom found good interrater reliability.[2] Caregiver responses that matched CBCL items and youth responses that matched YSR items were then automatically categorize according to CBCL/YSR narrowband, broadband, and DSM-oriented scales (Achenbach & Rescorla, 2001). For example, a caregiver-identified top problem of "fights with others" would be coded as a match to CBCL item 37, "Gets in many fights," and thus categorized as a fit to the Aggressive Behavior narrowband scale, the Externalizing Problems broadband scale, and the Conduct Problems DSM scale. For problems not matching any CBCL/YSR item, 21 additional codes were used (e.g., youth concerns about parents' divorce, parent concerns about youths' personal hygiene).

To assess interrater reliability, two clinical psychology graduate students independently coded the data from 20 randomly-selected child participants and their caregivers. Kappa coefficients were calculated for individual item coding and for narrowband, broadband, and DSM-oriented scale assignments. Mean kappa coefficients were calculated by averaging kappas computed for

each of the three top problems. For youth problem codes, mean κ was .91. For parent-reported problem codes, mean κ was .81. For the YSR, mean κ was .87 for narrowband categorizations, .83 for broadband categorizations, .and .91 for DSM scale categorizations. For the CBCL, mean κ was .78 for narrowband, .84 for broadband, and .85 for DSM scale classification. Thus, interrater reliability was strong for the coding system.

For psychometric analyses (below), TP Internalizing, Externalizing and Total scores were created, using the 0-10 ratings given to each problem by youths and caregivers. A TP Internalizing score was created for each youth and each caregiver at each assessment point by calculating the mean of ratings for problems coded as mapping onto the YSR/CBCL Internalizing broadband scale. A parallel procedure was used to create a TP Externalizing score. If no top problems fit the YSR/CBCL Internalizing broadband scale, the youth or caregiver's TP Internalizing score was 0; if no top problems fit CBCL/YSR Externalizing, the youth or caregiver's Externalizing score was 0. In addition, a TP Total score was created; this was the mean of all three top problems ratings regardless of how the problems had been categorized.[3]

Results

Top Problem Characteristics and Rating Patterns

Results of the top problem coding showed that 95.7% of the caregiver-identified problems matched a CBCL item, and 97.9% of the youth-identified problems matched a YSR item, suggesting that nearly all problems identified were significant and clinically relevant. Youth and caregiver wording of the problems generally added significant detail beyond that captured by CBCL/YSR items. For example, CBCL/YSR item 103 is "Unhappy, sad, depressed;" examples of top problems fitting that category included "Sad because she doesn't have a dad anymore," and "He is steady depressed, down all the time, even when doing something he likes." CBCL

and YSR item 112 is "Worries;" examples of top problems fitting that category included "Worried about parents being in the car and worrying something bad might happen," and "Worries about dying when he is 12." Of the caregiver top problems that matched a CBCL item, 94.9% fit a CBCL narrowband syndrome scale, and 72.2% fit the broadband Internalizing or Externalizing scale. Of the youth top problems that matched a YSR item, 98.1% fit a YSR narrowband scale, and 74.0% fit a broadband scale.

To assess strength of association among the three problem ratings of youths and caregivers separately, pairwise correlations were calculated (problem 1 vs. 2, 2 vs. 3, and 1 vs. 3) at weeks 1, 7, and 11. These ranged from .41 to .66 for caregivers and .31 to .61 for youths (all ps < .01). Youth-caregiver agreement was also assessed, by randomly selecting 20% of the youth-caregiver pairs and calculating agreement based on CBCL/YSR item coding for all items that had a possible match (i.e., excluding CBCL items that had no YSR counterpart). The mean percentage of youth-identified top problems that matched any of their caregivers' top problems was 29.03% at the individual item level, 60.86% at the narrowband scale level, 79.19% at the broadband scale level, and 58.29% at the DSM scale level. The mean percent of caregiver top problems that matched any of their youths' top problems was 26.03% at the item level, 54.90% at the narrowband level, 73.99% at the broadband level, and 51.88% at the DSM scale level. Thus, the information provided by youths and caregivers showed some agreement but was clearly not redundant.

Does the Top Problems Measure Complement Standardized Assessment?

Standardized measures such as the CBCL and YSR serve an important nomothetic objective by locating youths within a common set of empirically-derived scales. Leaders in the field have stressed the need to use such measures to guide clinical intervention, noting that efforts to use

evidence-based practice will be incomplete unless EBTs are undergirded by evidence-based assessment (see e.g., Achenbach, 2005; Hunsley & Mash, 2007). A question for the present study was whether—for those who adopt this goal of using empirically sound measurement to guide practice—assessing youth- and caregiver-identified top problems might add usefully to such standardized measures as the CBCL and YSR. That is, might top problems assessment add specific treatment-relevant information that is not so readily available through the standardized measures? At the individual problem level, the answer to this question was self-evident from the design of the measures; the CBCL and YSR are designed to show which problems on a standard list are rated "very true or often true" but not to determine which problems are of greatest concern to caregiver or child; TP assessment, by design, serves that prioritizing function and (see above) may also add specificity (e.g., what the youth is anxious about, what situations bring on sadness or depression).

At the CBCL/YSR scale level, a key question is whether identifying top problems provides information not yielded by standard use of the scales. For clinicians who want their practice to be informed by such evidence-based assessment as the CBCL and YSR (see Achenbach, 2005; Hunsley & Mash, 2007), a standard procedure that is used to help identify treatment targets involves noting which scales have scores in the clinical range; we assessed the extent to which using the scales in this way might miss problems caregivers and youths identify as most important to them. Note that this is a different question than the one addressed in the first paragraph of the Results section. That first paragraph noted that more than 95% of the top problems identified by caregivers and youths corresponded to items of the CBCL/YSR, and that most of the problems that matched a CBCL/YSR item also fit a CBCL/YSR narrowband and broadband scale. We now focus on the question of whether the top problems identified concerns

that did *not* show up *as items on clinical range scales* of the CBCL/YSR (and thus would not have been highlighted for attention via standard use of the CBCL/YSR). This analysis involved calculating the extent to which the problems identified in the TP assessment did *not* correspond to items on those CBCL/YSR narrowband and DSM scales *that were in the clinical range* (i.e., T=70 or higher). The analysis showed that for those cases where there was at least one clinical-range narrowband T-score on the CBCL, 43% of the caregiver top problems did not correspond to any item on any clinical range scale. Likewise, in those cases having at least one clinical-range narrowband T-score on the YSR, 69% of the youth top problems did not match any item on any clinical range scale. Corresponding figures for the CBCL/YSR DSM scales were 46% for caregiver-identified problems and 71% for youth-identified problems.

The analyses just reported focused on the percentage of *problems* identified through TP assessment that did not correspond to clinical range scales of the CBCL and YSR. The next set of analyses focuses on the percentage of *caregivers and youths* whose top problems were not evident in any clinical-range scales of the CBCL/YSR. Included as non-matches were those instances in which a caregiver or youth had no scale score in the clinical range. This procedure showed that for 41% of caregivers and 79% of youths no identified top problem corresponded to any item of any clinical range narrowband scale. Similarly, for the DSM scales, for 38% of caregivers and 80% of youths no identified top problem matched any item of any clinical range scale. These findings indicated that the TP measure added to the information generated by standard application of widely-used empirically-based standardized measures for caregivers and youths.

Top Problems Measure Psychometrics: Test-retest Reliability

To assess test-retest reliability of the TP Internalizing, Externalizing, and Total scores and to check against drift over time, correlations were examined for three pairs of time points: Calls 1 and 2, 7 and 8, and 11 and 12. The target inter-call interval was 7 days, but actual intervals varied (e.g., when multiple attempts were required to reach a family); the test-retest analyses only included calls occurring 5-21 days apart. The mean interval for youths for was 8.73 days (SD=3.11) for calls 1 and 2, 8.52 (SD=3.20) for calls 7 and 8, and 8.48 (SD=2.88) for calls 11 and 12. Mean caregiver interval was 8.42 (SD=2.67) for calls 1 and 2, 8.38 (SD=2.94) for calls 7 and 8, and 8.61 (SD = 3.36) for calls 11 and 12. As Table 1 shows, correlations were uniformly high, ranging from .69 to .91, all significant at $p<.01$.

Top Problems Measure Psychometrics: Convergent Validity

Convergent validity was assessed by computing correlations between caregiver and youth TP Internalizing scores from the first phone call and CBCL and YSR Internalizing broadband scales, narrowband scales, and DSM internalizing disorder scales from the initial assessment; correlations were also computed between caregiver and youth TP Externalizing scores and CBCL and YSR Externalizing broadband, narrowband, and DSM externalizing disorder scales. The youth TP Internalizing score was significantly correlated with YSR Internalizing and with the relevant YSR narrowband and DSM scales (Table 2). The youth TP Externalizing score was significantly correlated with YSR Externalizing and with the relevant YSR narrowband and DSM scales (Table 2). The caregiver TP Internalizing score was significantly correlated with CBCL Internalizing and with the relevant CBCL narrowband and DSM scales (Table 2). And the caregiver TP Externalizing score was significantly correlated with CBCL Externalizing and with the relevant CBCL narrowband and DSM scales (Table 2). Convergent validity was also assessed for TP Total. Youth TP Total was correlated .25 with YSR Total and .34 with BPC

Total, and caregiver TP Total was correlated .32 with CBCL Total and .48 with BPC Total; all these coefficients showed significant (p<.01) convergence between TP Total and Total Problem scores on the standardized measures.

Top Problems Measure Psychometrics: Discriminant Validity

Discriminant validity was assessed by examining correlations between TP Internalizing and Externalizing scores and the non-corresponding CBCL and YSR scores—e.g., TP Internalizing vs. CBCL/YSR Externalizing. The findings appear in parentheses in Table 2. Correlations between TP scores and measures of theoretically distinct constructs were small and generally nonsignificant. One exception was that caregiver TP Externalizing was significantly correlated -.24 with the DSM Anxiety scale, suggesting an inverse association rather than a negligible one. It was hard to identify an ideal target measure for discriminant validity assessment in relation to caregiver TP Total, because all the caregiver measures were problem reports and thus likely to be positively correlated with other measures of total problems, including TP Total. So, the analyses employed the Brief Symptom Inventory, completed by caregivers in relation to their own functioning; this held reporter constant but not assessment target. The BSI correlated .14 (ns) with caregiver TP Total, suggesting that caregiver TP Total was specific to the assessment target and not reflective of general reporting style. The Positive Qualities Scale of the YSR was used as a discriminant criterion for youth TP Total, and these scales correlated -.08 (ns).

Top Problems Measure Psychometrics: Change Over Time in Top Problems Scores

Criterion validity was assessed by testing whether TP scores declined over time. [Youths in treatment generally show symptom reduction over time (see Weisz, 2004).] Slopes and intercepts were calculated with random coefficient growth models using the Hierarchical Linear Modeling Platform (HLM 6.08; Raudenbush, Bryk & Congdon, 2004). Time was used as the Level 1

predictor and participant (youth) as the Level 2 grouping variable.[4] Intraclass correlation coefficients (ICCs) were examined to assess the degree to which youth and caregiver TP scores clustered within therapists. ICCs were small for both youth and caregiver ratings (range: .0002 to .014), so a therapist level was not included in the models. The actual spacing of the calls was used, with a natural logarithmic transformation applied to linearize curvilinear time trends (Cohen, Cohen, West, & Aiken, 2003). Analyses also examined slope reliability, the ratio of true to observed scores; in random coefficient models, the reliability of each coefficient is the ratio of parameter variance to total variance (parameter variance + error variance). Slope reliabilities are thus a function of (1) the degree to which slopes differ across individuals, and (2) the precision with which each individual's slope is estimated (Raudenbush & Bryk, 2002).

Separate growth models were calculated for youth and caregiver reports, using Top Problems 1, 2, 3, and TP Internalizing, Externalizing, and Total. Models were also estimated for combined (youth + caregiver) TP scores. To ensure sufficient data for fair slope estimates, the analyses used observations through 26 phone calls; the sample in later calls fell below 50%, as youths completed treatment. A total of 13 unconditional growth models were examined to assess change over time in TP scores.

Slope reliability. Reliabilities for the slopes of Top Problems 1, 2, 3, and TP Internalizing, TP Externalizing, and TP Total regressed on log of days were quite similar to slope reliabilities for the corresponding BPC scales. Individual TP reliabilities ranged from .72 to .80 for caregivers and .70 to .77 for youths, Internalizing, externalizing, and total score reliabilities ranged from .75 to .81 for caregiver TP and .71 to .80 for caregiver BPC; corresponding reliability ranges were .78 to .83 for youth TP and .82 to .85 for youth BPC. These values indicate that repeated administration of the TP measure provided reliable estimates of change over time.

Criterion Validity--Change over Time. All models for both caregivers and youths examining change in Top Problems 1, 2, 3, and TP Internalizing, Externalizing and Total yielded significant ($p<.01$) positive intercepts and significant negative slopes ($p<.01$). This indicates that all severity ratings were significantly different from zero at the outset, as would be expected for youths entering treatment, and that the severity of TP 1, 2, 3, Externalizing, Internalizing, and Total declined significantly over time as rated by both caregivers and youths, as would also be expected for youths over time in treatment.

Convergent validity over time. Two approaches were used to assess convergent validity over time: Parallel process growth models and correlating the empirical Bayes slope estimates for TP Internalizing, Externalizing, and Total for youths and caregivers with corresponding BPC scales. We fit parallel process growth models in Mplus Version 6.00 (Muthén & Muthén, 1998-2010). For these analyses we fit models using Maximum Likelihood estimation (missing data were estimated using Full Information Maximum Likelihood), in which the latent intercepts and slopes for TP Internalizing, Externalizing, and Total were included in the same model with latent intercepts and slopes for the BPC Internalizing, Externalizing, and Total, respectively, for youths and caregivers. Time scores in the models were allowed to vary within and across respondents. Two models were compared, one with the slopes correlated and one with slopes uncorrelated, and the Bayesian Information Criterion (BIC) was chosen as an indicator of model fit because of its relative resistance to the influence of changes in numbers of parameters on model fit. Smaller BIC values indicate better fit (Rafferty, 1995).

In all cases, models specifying correlated slopes and intercepts had BIC values superior to those specifying uncorrelated slopes and intercepts. In addition, the covariances of the latent slopes and intercepts for measures in the same models were all significantly related at $p<.01$.

Empirical Bayes estimates (Table 3) showed that the standardized correlations between corresponding slopes were medium to large (Cohen, 1992) and were all significant. The patterns of change over time in TP and BPC measures are shown in Figures 1 (caregivers) and 2 (youths).

Discussion

Leaders in treatment research have often called for assessments that rely on the perspective of clients, and some have documented clinician concerns about the utility or sufficiency of the standardized outcome measures often used in research (e.g. Bickman et al, 2000; Garland, Kruse, & Aarons, 2003); clinician suggestions about how to improve the utility of assessment tools have stressed the need for brevity and simplicity of measurement (see e.g., Garland et al., 2003). The present study explored whether a type of client-guided assessment that is widely-used in clinical practice might be structured to form a brief, simple, psychometrically sound idiographic complement to standardized assessment of mental health symptoms. The findings suggest that the resulting approach—having clients identify and repeatedly rate the severity of their top problems—may contribute to both clinical practice and treatment research in a number of ways. The problems identified by youths and caregivers were clinically significant concerns; more than 95% of both youth and caregiver problems were reliably coded as matches to CBCL and YSR items. To assess the potential contribution of top problem assessment, we investigated (a) whether the top problems identified by youths and caregivers adds to the information standardized measures provide, and (b) whether the TP measure provides psychometrically sound assessment.

As to the first question, the specificity of youth and caregiver top problems complemented the more general information provided by standardized measures. The CBCL/YSR provided a rich array of valuable data, in the form of 0-1-2 ratings on problem items, and scores on narrowband

and DSM scales, with scale scores in the clinical range warranting clinical attention. Such information provides an excellent index of severity relative to normative samples, and multiple scales that can be rank-ordered in terms of their statistical deviance; but that information is not designed to reveal which specific problems represent the highest priority treatment targets for caregiver or youth. TP assessment complements the CBCL and YSR by providing that information on client problem priorities. TP assessment also adds useful specificity to CBCL and YSR information, as when item 112 on the CBCL or YSR shows that a youth "worries," but top problem identification shows *what* the youth worries about (e.g., in our sample, dying, falling into a septic tank, "losing my house and family," and "something bad happening to Mom").

As further evidence on how TP assessment complements standardized assessment, our analyses showed that the standard approach of using the CBCL and/or YSR to identify treatment targets by identifying scale scores in the clinical range, though clinically appropriate and valuable, would often miss the problems that are most important to youths and caregivers. For example, for 41% of caregivers and 79% of youths none of their identified top problems corresponded to any clinical range CBCL/YSR narrowband scale; and for 38% of caregivers and 80% of youths no identified top problem matched any of the clinical range DSM scales. Our findings thus suggest that the TP measure adds importantly to the information available through the CBCL and YSR, arguably the most widely-used and thoroughly-researched empirically-based standardized measures for caregivers and youths. Our findings suggest that a combination of standardized and idiographic assessment may help ensure attention to both scientifically-derived dimensions of psychopathology and the specific problems of greatest concern to clients.

Our findings also supported the psychometric strength of the TP measure. The evidence supported (a) test-retest reliability, assessed at three different times in treatment; (b) convergent

validity vis-à-vis theoretically-related dimensions of psychopathology assessed via standardized measures; (c) discriminant validity in relation to theoretically distinct dimensions of psychopathology assessed via standardized measures; (d) sensitivity to change over time; (e) slope reliability; and (f) significant associations between slopes for the TP measures and slopes generated by a psychometrically sound criterion measure of clinical change (i.e., the BPC). Taken together, the findings suggest that the TP measure has the psychometric strength needed to make it a viable tool for clinical research and practice.

The feasibility of including the TP measure in treatment research is enhanced by the brevity of our approach and the ease of repeated administration to caregivers and youths over time. If the TP measure were combined with an efficient standardized measure (e.g., the BPC), the combination could rather easily be included in weekly assessments throughout treatment. Such weekly tracking would provide the kind of detailed documentation of trajectories across episodes of care that is needed for multilevel modeling and related approaches to analysis of group differences in treatment impact. This would be helpful, for example, in effectiveness trials in which manual-guided treatments of standard duration are compared to usual care of unlimited duration; in these cases, post-treatment assessment alone is difficult to interpret given duration and dose differences, and trajectories of change provide more unbiased estimates of group differences in treatment impact.

In clinical practice, such frequent monitoring might be done by staff who call clients to confirm appointments each week, or by clients in a waiting room before each session begins. Week-by-week trajectories of change—on individual problems and on the total score for youth-report and caregiver-report problems—could be charted, graphed, and monitored for each case, thus informing the therapist and supervisor of the extent to which the top problems that led to

treatment are being successfully addressed. This information could contribute significantly to ongoing treatment planning and clinical supervision, with intervention strategies adjusted for those problems that are not showing reduced severity over time in treatment. For practitioners interested in gauging the success of their interventions, weekly TP assessment could provide consumer-sensitive information, plus protection against the loss of outcome information when clients end treatment by simply not showing up. In such cases, the last weekly top problem ratings would mark the end of treatment, and the full picture of week-to-week change throughout the episode would immediately be available to the clinician. The approach might also enhance clinician awareness of differences between caregivers and youths in the ways they conceptualize and prioritize problems; this in turn could help clinicians respond appropriately to the differing needs and goals of caregivers and youths in treatment, and thus potentially improve rapport, treatment engagement, retention, and outcome.

The study had certain limitations that warrant attention. First, the TP measure relies on subjective reports by youths and caregivers; there is no known gold standard validity criterion that can be used to check on whether the problems identified in our interviews were in fact the most important in some objective sense. Second, because phone calls were used for assessment, the study could not tell us whether the psychometric findings would have been different with a different assessment format or context. Third, because our focus was on change over time in severity ratings of the problems identified at the beginning of treatment, the study did not provide information on other problems that might have arisen after treatment began. Our focus is consistent with a model in which a treatment "contract" is established at the outset, by identifying the problems that are to be addressed, and success in fulfilling that contract is monitored by tracking change in those problems. To the extent that treatment is associated with

reductions in severity of the initial problems, one would expect concern about those problems to diminish, and of course new problems might well arise since treatment takes place in the context of everyday life. Although the emergence of new problems was not our focus, and would not affect the validity or utility of our findings, the emergence of new problems during treatment could be an interesting focus for future research. Indeed, such research might use assessment methods like ours at the outset of treatment, but repeat the assessment periodically thereafter. Finally, although our sample was ethnically diverse, the sample available for any specific ethnic group was too small to permit meaningful subgroup analyses of ethnic group differences in measure characteristics; this might certainly be a useful focus for future study.

As for the TP measure itself, both limitations and strengths can be noted. Problems were assessed immediately after diagnostic assessment, with participants listing their main concerns, rating them, then identifying the top three; and, to keep the focus on serious, clinically relevant concerns, youths and caregivers were asked to identify problems that were current, giving them trouble, and important to work on. This approach (i.e., after diagnostic assessment) seemed efficient and ecologically valid (diagnostic assessment in some form is required for reimbursement in most practice settings). However, a more streamlined procedure might also be used in clinical practice; future work could test whether simpler methods (e.g., simply asking youths and caregivers to identify and rate their top three problems) would suffice, or would lead to less clinically relevant identified problems. Our focus on only three top problems may limit attention to other concerns that matter to clients; emphasizing the top three supports efficiency and clinical focus, but expanded versions might be tried in future research. Obtaining top problem ratings by phone may not be ideal in all cases, particularly for clients who find phone interviews aversive, or for those who are difficult to reach by phone. Alternative approaches

could certainly be used to collect these simple ratings. In addition to these potential limitations, the TP measure also offers several strengths. One is that asking clients to report on their top problems is ecologically valid as an efficient, respectful, clinically-sensitive approach that builds directly on procedures already common in clinical practice. The repeated assessment approach is quite accessible to clinicians, given the brevity and simplicity of the TP measure. Finally, the rating system is simultaneously idiographic and systematic, thus having the potential to support clinical practice (e.g., monitoring treatment response and progress toward goals) and clinical research (e.g., measuring trajectories of change), as described in the introduction.

Broadly construed, the study findings may support advances in assessment that could improve both clinical research and clinical practice. On the research front, the TP measure may provide a way to complement the study of empirically-derived constructs with attention to client-derived problems without compromising psychometric integrity. On the clinical practice front, the efficient assessment approach described here may provide a way to engage some of those practitioners who report that they do not value or use systematic outcome assessments to identify treatment foci or monitor treatment progress (Bickman et al., 2000; Garland et al., 2003; Hatfield & Ogles, 2004). If this kind of evidence has clinical appeal, perhaps it can contribute to making everyday clinical care more systematic, evidence-informed, and effective.

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

Bannon, W.M., & McKay, M.M. (2005). Are barriers to service and parental preference match for service related to urban child mental health service use? *Families in Society: The Journal of Contemporary Social Services, 86,* 30-34.

Bickman, L., Rosoff, J., Salzer, M.S., Summerfelt, W.T., Noser, K., Wilson, S.J., & Karver, M.S. (2000). What information do clinicians value for monitoring adolescent client progress? *Professional Psychology: Research and Practice, 31,* 70-74.

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.

Chorpita, B.F., Reise, S., Weisz, J.R., Grubbs, K., Becker, K.D., & Krull, J.L. (2010). Evaluation of the Brief Problem Checklist: Child and Caregiver Interviews to Measure Clinical Progress. *Journal of Consulting and Clinical Psychology, 78,* 526-536.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cone, J.D. (1999). Introduction to the special section on self monitoring: A major assessment method in clinical psychology. *Psychological Assessment, 11,* 401-411.

Crane, D.R., Griffin, W., & Hill, R.D. (1986). Influence of therapist skills on client perceptions of marriage and family therapy outcome: Implications for supervision. *Journal of Marital and Family Therapy, 12,* 91-96.

Cytrynbaum, S., Ginath, Y., Birdwell, J., & Brandt, L. (1979). Goal attainment scaling: A critical review. *Evaluation Quarterly, 3,* 5-40.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13*, 595-605.

Doss, B.D., Simpson, L.E., & Christensen, A. (2004). Why do couples seek marital therapy? *Professional Psychology: Research and Practice, 35,* 608-614.

Doss, B.D., Thum, Y.M., Sevier, M., Atkins, D.C., & Christensen, A. (2005). Improving relationships: Mechanisms of change in couple therapy. *Journal of Consulting and Clinical Psychology, 73,* 624-633.

Ebesutani, C., Bernstein, A., Martinez, J. I., Chorpita, B. F., & Weisz, J. R. (in press). The Youth Self Report: Applicability and validity across younger and older youths. *Journal of Clinical Child and Adolescent Psychology*

Eifert, G.H., Evans, I.M., & McKendrick, V.G. (1990). Matching treatments to client problems not diagnostic labels: A case for paradigmatic behavior therapy. *Journal of Behavior Therapy and Experimental Psychiatry, 21,* 163-172.

Garland, A., Kruse, M., & Aarons, G.A. (2003). Clinicians and outcome measurement: What's the use? *Journal of Behavioral Health Services and Research, 30*, 393-405.

Goodyer, I.M., Herbert, J., Secher, S.M., & Pearson, J. (1997). Short-term outcomes of major depression: I. comorbidity and severity at presentation as predictors of persistent disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 179-187.

Hafkenscheid, A. (1991). Psychometric evaluation of a standardized and expanded Brief Psychiatric Rating Scale. *Acta Psychiatria Scandinavia, 84,* 294-300.

Hatfield, D.R., & Ogles, B.M. (2004). The use of outcome measures by psychologists in clinical practice. *Professional Psychology: Research and Practice*, *35*, 485–491

Hawley, K.M., & Weisz, J.R. (2003). Child, parent, and therapist (dis)agreement on target problems in outpatient therapy: The therapist's dilemma and its implications. *Journal of Consulting and Clinical Psychology, 71*, 62-70.

Haynes, S. N., & O'Brien. W. O. (2000). *Principles and practice of behavioral assessment.* New York, NY: Plenum/Kluwer Press.

Hoagwood, K.E., Cavaleri, M.A., Olin, S.S., Burns, B.J., Slaton, E., Gurttadaro, D., Hughes, R. (2010). Family support in children's mental health: A review and synthesis. *Clinical Child and Family Psychology Review, 13*, 1-45.

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.

Kazdin, A.E. (2000). *Behavior modification in applied settings* (6th ed.). New York: Wadsworth.

Kazdin, A.E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3,* 1-27.

Kiresuk, T.J., & Sherman, R.E. (1968). Goal attainment scaling: a general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal, 4,* 443-453.

Kolko, D., & Kazdin, A.E. (1993). Emotional/behavioral problems in clinic and nonclinic children: Correspondence among child, parent, and teacher reports. *Journal of Child Psychology and Psychiatry, 34*, 991-1006.

Kovacs, M., Obrosky, D.S., & Sherrill, J. (2003). Developmental changes in the phenomenology of depression in girls compared to boys from childhood onward. *Journal of Affective Disorders, 74,* 33-48.

Lambert, M.C., Rowan, G.T., Lyubansky, M., & Russ, C.M. (2002). Do problems of clinic-referred African-American children overlap with the Child Behavior Checklist? *Journal of Child and Family Studies, 11,* 271-285.

Lambert, M.J., 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.

Lambert, M. J., Shapiro, D.A., & Bergin, A. E. (1986). The effects of psychotherapy. In S.L. Garfield & A. E. Bergin (Eds.). *Handbook of Psychotherapy and Behavior Change*. (3rd ed.) New York: John Wiley and Sons.

MacKay, G., Somerville, W., & Lundie, J. (1996). Reflections on goal attainment scaling (GAS): cautionary notes and proposals for development. *Educational Research, 38,* 161-172.

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

Muthén, L.K. and Muthén, B.O. (1998-2010). Mplus users guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Ollendick, T.H., & King, N.J. (1994). Diagnosis, assessment, and treatment of internalizing problems in children. *Journal of Consulting and Clinical Psychology, 62,* 918-927.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. In A. E. Raftery (Ed.), *Sociological Methodology 1995* (pp. 111-164). Oxford: Blackwell.

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

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Rescorla, L., et al. (2007). Epidemiological comparisons of problems and positive qualities reported by adolescents in 24 countries. *Journal of Consulting and Clinical Psychology, 75,* 351-358.

Rooney, M. T., Fristad, M. A., Weller, E. B., & Weller, R. A. (1999). *Administration manual for the Children's Interview for Psychiatric Syndromes (ChIPS).* Washington, D.C.: American Psychiatric Press, Inc.

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

Weersing, V. R., & Weisz, J. R. (2002). Mechanisms of action in youth psychotherapy. *Journal of Child Psychology and Psychiatry, 43*, 3-29.

Weisz, J.R. (2004). *Psychotherapy for children and adolescents: Evidence-based treatments and case examples.* Cambridge, UK: Cambridge University Press.

Weisz, J.R., & Chorpita, B.C. (in press). Mod squad for youth psychotherapy: Restructuring evidence-based treatment for clinical practice. In P.C. Kendall (Ed.), *Child and adolescent therapy: Cognitive-behavioral procedures, 4th edition.* New York: Guilford.

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., et al. (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.

Weisz, J.R., & Weiss, B. (1991). Studying the 'referability' of child clinical problems. *Journal of Consulting and Clinical Psychology, 59*, 266-273.

Weller, E., Weller, R., Fristad, M., Rooney, M. & Schecter, J. (2000). Children's interview for psychiatric syndromes (ChIPS). *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 76-84.

Yeh, M., & Weisz, J.R. (2001). Why are we here at the clinic? Parent-child (dis)agreement on referral problems at treatment entry. *Journal of Consulting and Clinical Psychology, 69*, 1018-1025.

Author Note

Address correspondence to John R. Weisz, Department of Psychology, William James Hall, Harvard University, 33 Kirkland Street, Cambridge, MA 02138 (jweisz@jbcc.harvard.edu).

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 included: John Weisz, (Network Director), Bruce Chorpita, Robert Gibbons, Charles Glisson, Evelyn Polk Green, Kimberly Hoagwood, Kelly Kelleher, John Landsverk, Stephen Mayberg, Jeanne Miranda, Lawrence Palinkas, Sonja Schoenwald.

The research activities and study personnel were supported by funding to John Weisz from the Norlien Foundation, the John D. and Catherine T. MacArthur Foundation, and the National Institute of Mental Health (MH068806, MH085963), by funding to Bruce Chorpita from the John D. and Catherine T. MacArthur Foundation and the Annie E. Casey Foundation, and by funding to Sarah Kate Bearman from the National Institute of Mental Health (MH083887). We are grateful to the clinical service organizations, clinicians, youths, and parents who participated in the project.

Footnotes

[1]Because our sample extended down to age 7, it should be noted that reliability and validity of the YSR with ages 7-10 have been supported in multiple studies; YSRs completed by 7-10 year-olds have been found to be very similar to YSRs of older children in (a) internal consistency and test-retest reliability of Internalizing, Externalizing, and Total Problems scores (Yeh & Weisz, 2001); (b) parent-child and teacher-child agreement on Internalizing, Externalizing, and Total Problems (Kolko & Kazdin, 1993); and (c) factor structure and strength of association of Internalizing, Externalizing, and Total Problems scores with multiple convergent and discriminant validity criteria (Ebesutani, Bernstein, Martinez, Chorpita, & Weisz, in press).

[2]This coding was done for the purpose of assessing measure psychometrics; such coding would not be expected to take place in the course of everyday clinical use of the Top Problems assessment approach. However, interested clinicians or researchers are welcome to use our coding manual, which can be obtained from the first author.

[3]Although gender and ethnicity were not primary foci of this study, some interesting group differences emerged in TP Total scores on the initial phone call. Girls rated themselves higher in severity than boys (means: 6.23 vs. 5.31), t (175) =2.08, $p<.05$, and ethnic minority youths rated themselves higher in severity than European-American youths (means: 5.98 vs. 5.10), $t(175) =2.20$, $p<.05$. In contrast, caregivers did not differ significantly in their TP Total ratings as a function of youth gender or ethnic group.

[4]Random effects were supported for slopes and intercepts for all level one models, indicating significant (i.e., $p < .001$) interindividual variation across the slopes and intercepts of all of our models.

Table 1

Top Problems Internalizing, Externalizing, and Total Test-Retest Reliability for Calls 1-2, 7-8 and 11-12: Caregiver and Youth Reports.

	Caregiver Report			Youth Report		
Call Number	TP Int	TP Ext	TP Total	TP Int	TP Ext	TP Total
Call 1-2	.86	.90	.77	.79	.80	.69
(*n*)	(167)	(167)	(167)	(167)	(167)	(167)
Call 7-8	.90	.89	.82	.91	.85	.88
(*n*)	(150)	(150)	(150)	(150)	(150)	(150)
Call 11-12	.91	.90	.89	.85	.81	.78
(*n*)	(135)	(135)	(135)	(136)	(136)	(136)

Note: TP = Top Problems; Int = Internalizing; Ext = Externalizing; All correlations are significant at $p < .01$.

Table 2

Correlations of Youth-Report Top Problems Internalizing and Externalizing Scales with Broadband, Narrowband, and DSM-Oriented Scales of the Youth Self-Report (N=133)

YSR/CBCL Scale[a]	Youth TP Int	Youth TP Ext	Caregiver TP Int	Caregiver TP Ext
Internalizing	.33**	(.13)	.38**	(-.06)
Externalizing	(.06)	.37**	(-.12)	.59**
Anxious Depressed	.33**	(.08)	.37**	(-.09)
Withdrawn-Depressed	.23*	(.09)	.29**	(.08)
Rule-Breaking	(.00)	.21*	(-.13)	.50**
Aggressive	(.09)	.42**	(-.10)	.56**
DSM Affective	.27*	(.08)	.27**	(.11)
DSM Anxiety	.39**	(-.04)	.37**	(-.24**)
DSM Oppositional	(.05)	.47**	(-.13)	.57**
DSM Conduct	(.02)	.31**	(-.17*)	.56**

Note. Cohen (1992) suggests benchmarks of .10, .30, and .50 for small, medium, and large effects. Correlations in parentheses represent discriminant validity coefficients.

[a]The Youth TP Int and TP Ext columns show correlations of Youth Top Problems Internalizing and Externalizing with scale scores of the Youth Self-Report. The Caregiver TP Int and TP Ext columns show correlations of Caregiver Top Problems Internalizing and Externalizing with scale scores of the Child Behavior Checklist.

* $p<.05$

** $p<.01$

Table 3

Slope Correlations of Top Problems Internalizing, Externalizing, and Total with BPC Internalizing, Externalizing, and Total for Youth, Caregiver, and Combined

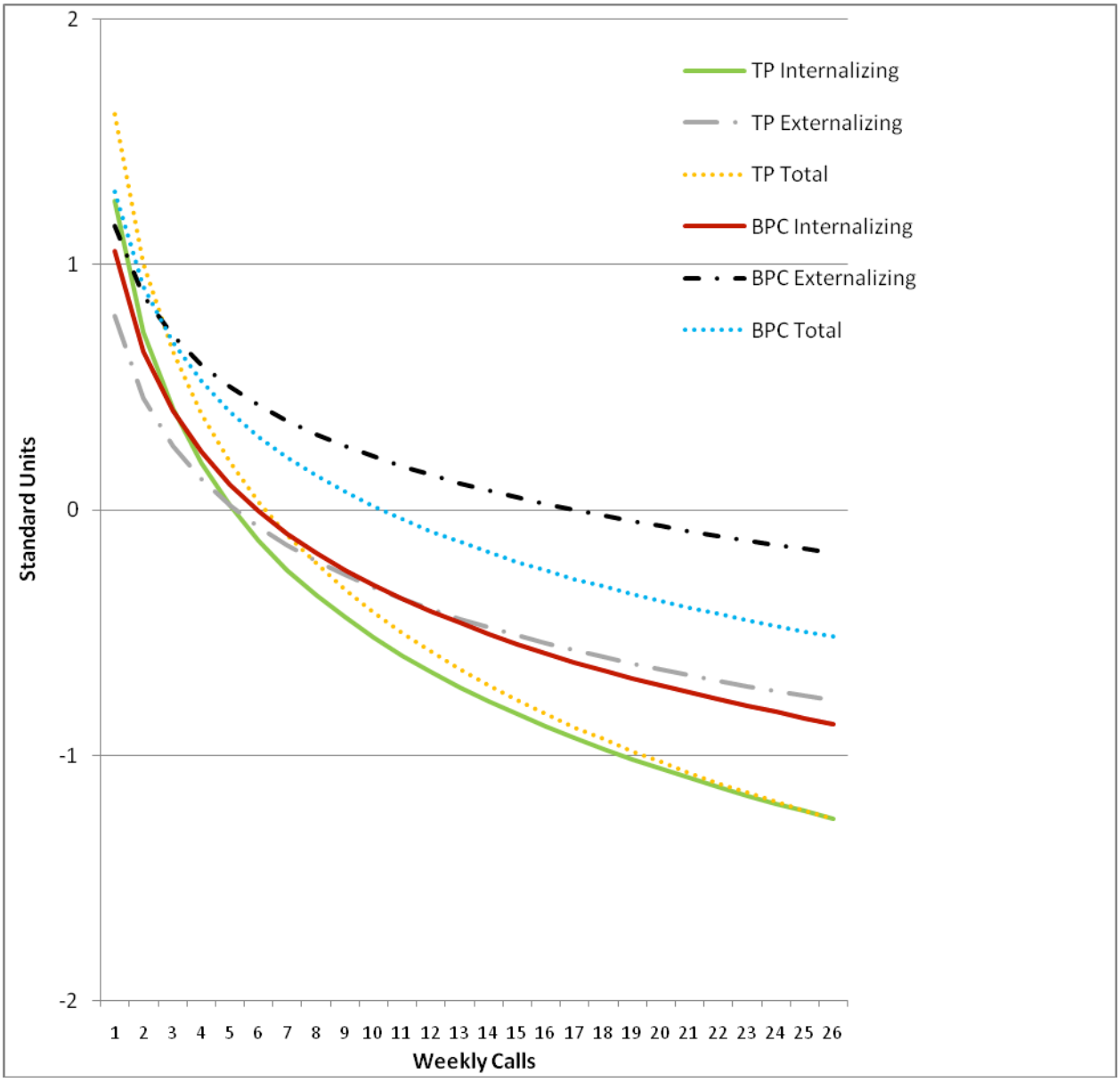
Reporter	Internalizing	Externalizing	Total
Youth	.27	.51	.44
Caregiver	.50	.54	.57
Combined (Youth + Caregiver)	.43	.61	.56

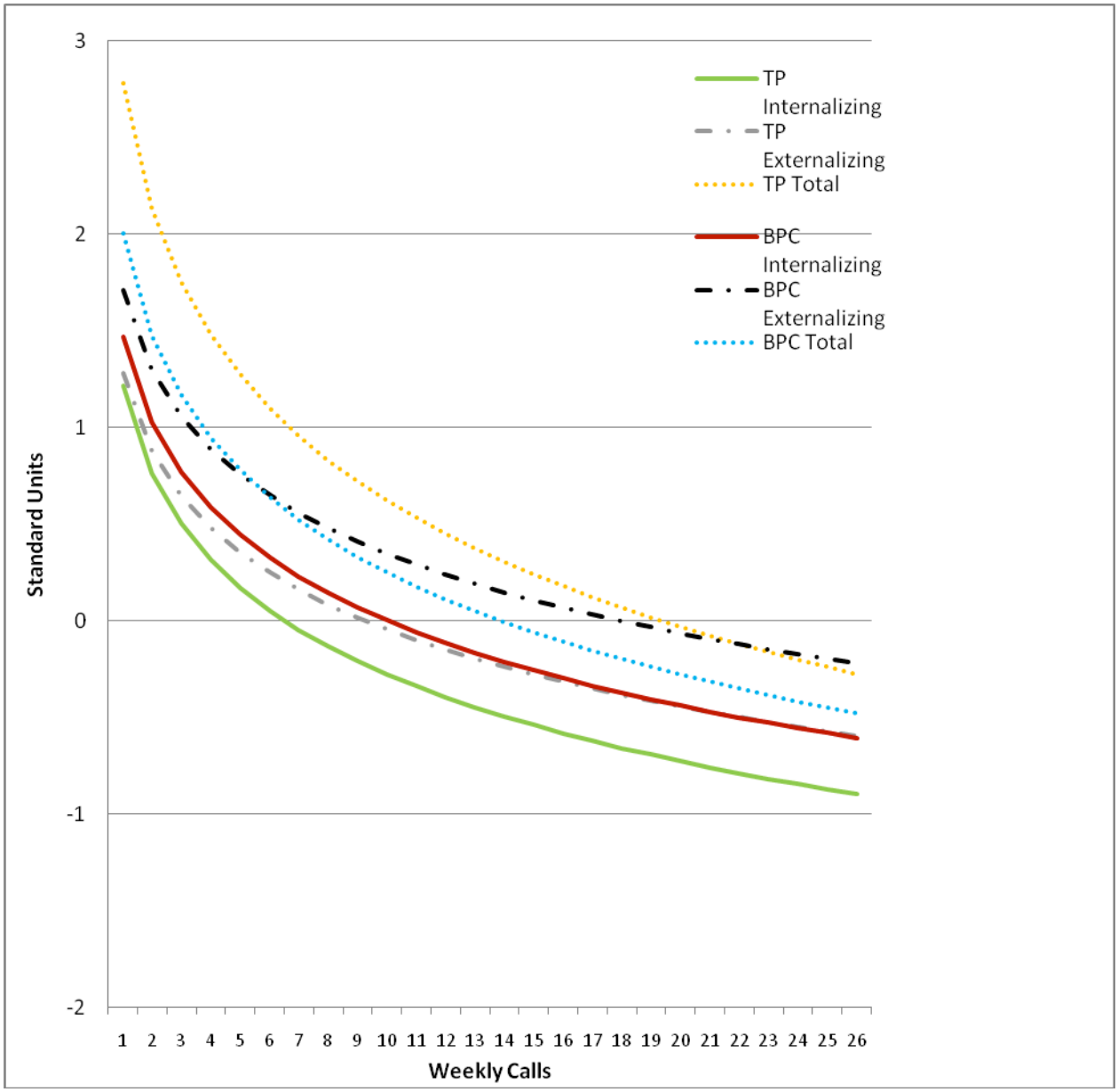
Note: All coefficients are significant at $p<.01$. Cohen (1992) suggests benchmarks of .10, .30, and .50 for small, medium, and large effects.

Figure Captions

Figure 1. Change Over Time in Top Problems and Brief Problem Checklist Standardized Scores by Caregiver Report

Figure 2. Change Over Time in Top Problems and Brief Problem Checklist Standardized Scores by Youth Report


2
1
0
-1
-2
Standard Units
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Weekly Calls
TP Internalizing
TP Externalizing
TP Total
BPC Internalizing
BPC Externalizing
BPC Total


3
2
1
0
-1
-2
Standard Units
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Weekly Calls
TP Internalizing
TP Externalizing
TP Total
BPC Internalizing
BPC Externalizing
BPC Total