Adm Policy Ment Health
DOI 10.1007/s10488-010-0332-x

ORIGINAL PAPER

Improvement in Symptoms Versus Functioning: How Do Our Best Treatments Measure Up?

Kimberly D. Becker · Bruce F. Chorpita · Eric L. Daleiden

Abstract We examined the effects of redefining standards of evidence for treatments targeting childhood mental health problems by expanding outcomes beyond symptom reduction to include functioning. Over 750 treatment protocols from 435 randomized controlled trials were rated based on empirical evidence. Nearly two-thirds (63.9%) demonstrated at least a minimum level of evidence for reducing symptoms; however, only 18.8% of treatments demonstrated evidence for reducing functional impairment. Of those treatments with empirical support for symptom reduction, the majority did not demonstrate empirical support for improvement in functioning because measures of functioning were not included in the studies in which these treatments were tested. However, even when measures of functioning were included, it was much more difficult for treatments to achieve improvement. Among treatments that achieved improvement in functioning, the most notable were Collaborative Problem Solving for disruptive behavior and Cognitive Behavioral Therapy plus Medication for traumatic stress because they demonstrated no support for symptom reduction but good support for improvement in functioning. Results are discussed within the context of evaluating the standards of evidence for treatments and the opportunity to move towards a multidimensional framework whose utility has the potential to exceed the sum of its parts.

Keywords Childhood treatments · Standards of evidence · Symptoms · Functioning

Significant advances have been made in the development and refinement of treatments aimed at improving children's mental and behavioral health. Concomitant with the proliferation of treatments over the last two decades has been the elaboration of frameworks for evaluating the efficacy of these interventions (cf. Chambless et al. 1996, 1998; Chorpita et al. 2002; Silverman and Hinshaw 2008; Task Force on Promotion and Dissemination of Psychological Procedures 1995). Across these evaluative frameworks, the efficacy of an intervention is determined by the degree to which it outperforms a comparison group on a specified outcome indicator within a randomized controlled trial (RCT).

Historically, the primary outcome domain of interest in treatment studies has been symptom reduction following treatment. The emphasis on symptom reduction has arisen as a natural result of treatment delivery under the tightly controlled conditions and methods of efficacy trials that, due to their concerns about internal validity, prioritize symptom measures to maximize construct specificity (Bickman et al. 2000; McGlinchey et al. 2008).

This manuscript was presented at the annual conference of the Association of Behavioral and Cognitive Therapies in November, 2010 in San Francisco, CA.

K. D. Becker (✉)
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 111 Market Place, Suite 901, Baltimore, MD 21205, USA
e-mail: kbecker@jhsph.edu

B. F. Chorpita
Department of Psychology, The University of California, UCLA, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563, USA
e-mail: chorpita@ucla.edu

E. L. Daleiden
PracticeWise, LLC, 285 Wilson Ave, Satellite Beach, FL 32937, USA
e-mail: eric@daleiden.com

Published online: 05 January 2011

The reliance on symptom reduction as the primary, if not sole, indicator of treatment efficacy has important implications for how we award the status of "evidence-based treatment" (EBT) to any given intervention. It is possible that our evaluative frameworks overestimate the strength of the evidence for our treatments in some cases, such as when symptoms improve following treatment but no improvement is demonstrated in other domains (e.g., functioning, family environment; Hoagwood et al. 1996; Kazdin and Kendall 1998). In other cases, it is possible that our criteria underestimate the strength of the evidence, such as when measures of other constructs (e.g., functioning, family environment) support the treatment, whereas indicators of symptom reduction do not (De Los Reyes and Kazdin 2008).

Given that the number of treatments designated as evidence-based due to their empirical support for reducing symptoms is in the hundreds and continues to grow (Chorpita et al. in press), perhaps the "evidence-based" label is no longer sufficiently informative to allow for the more demanding discrimination tasks we now face. It may be the case that the term "evidence-based" has become so ubiquitous in health care parlance that its utility as a treatment specifier has diminished, in part because it is relatively easy for treatments to demonstrate a relative advantage over control conditions with regard to symptom reduction (Wampold et al. 2009) and in part due to the lack of additional criteria to help guide selection among the EBTs currently on the list.

It is incumbent upon us to develop higher standards to allow for better discrimination among those treatments that earn the "evidence-based" designation. One solution is to consider new definitions of EBTs that involve a penalty function for weak or null findings, such that high-level treatments could potentially drop in status as the evidence evolves. Another approach is to consider how treatments fare with regard to multiple outcomes (e.g., symptom reduction, functioning, consumer satisfaction). Hoagwood and colleagues (Hoagwood et al. 1996) proposed a dynamic conceptual model for children's mental health treatment outcomes that encompasses five domains: symptoms, functioning, consumer perspectives, environments, and systems. They argued that the actual level of symptom severity is often less informative to the interpretation of a youth's outcome than is the context in which those symptoms are manifested (Hoagwood et al. 1996). In other words, a comprehensive assessment of treatment outcomes is necessary to truly understand the meaning of symptom change. Evaluation of treatments across multiple outcomes has the potential to refine our method of evaluation in a way that addresses the problem of proliferation of findings and yields goal-specific guidance in treatment identification and selection. Just as we often need to clarify that a treatment is evidence-based for a given problem or age group, it may be useful to specify the particular outcome domain for which a treatment is evidence-based. The task is then framed not simply as identifying the best practices, but the best practice for a particular purpose.

In particular, one might consider inclusion of improvement in functioning as an outcome criterion. Symptom presentation and functional impairment may be conceptualized as distinct outcomes. Symptom presentation could refer to the emotional or behavioral manifestations (e.g., avoidance, firesetting, difficulty sustaining attention, decreased appetite) of a particular problem (e.g., anxiety, conduct problems, depression, inattention). Consideration of the frequency, duration, and constellation of symptoms, therefore, could provide an index of the severity of a particular problem or disorder (Hoagwood et al. 1996). Functioning may be conceptualized as a continuum that reflects a youth's deficits in carrying out daily roles on the one end as well as a youth's ability to adapt to the demands of and carry out daily roles on the other (Hoagwood et al. 1996; Winters et al. 2005). Even beyond adaptive functioning on the continuum is positive functioning, which is conceptualized as a youth's strengths to exceed role expectations; in other words, thriving (Karver and Bickman 2002). Although symptom severity and functional impairment may be positively associated, this is association is not a foregone conclusion. It may be the case that a youth with relatively severe symptoms demonstrates relatively adept functioning, perhaps due to personal resilience, compensatory behaviors, or environmental supports (Bird 1999; Pickles et al. 2001). As an example, a high school student might engage in significant binge-drinking or cannabis use but demonstrate no decrements in school or sports performance, no interference with part-time employment, and no curfew violations. Likewise, it is possible that a youth with less severe or subclinical symptoms might experience significant functional impairment, or variable impairment across settings (Hoagwood et al. 1996; Pickles et al. 2001). Research also suggests that the slope of the linear relationship between symptom severity and functional impairment may vary according to problem area, with phobias and anxiety related to lower functional impairment than depression and externalizing disorders (Simonoff et al. 1997), although other research suggests that depression may be related to higher levels of functional impairment than conduct problems (Pickles et al. 2001).

Inclusion of functioning as an outcome indicator has the potential to improve our evaluative frameworks for EBTs not only by improving the scientific method of judging our treatments, but also by potentially enhancing the relevance of resulting evidence summaries for mental health stakeholders in a number of ways (Chorpita 2001; Rapp et al. 2005). First, it is often not the symptoms, but the disruption

or interference that creates hardships for youth and results in identification and treatment seeking (Bird et al. 1990). Improvement in functioning may be related to client views of treatment success (Zimmerman et al. 2006) and clinician views of client needs (Striley et al. 2003; Hodges et al. 2000). Functioning also provides a context for the interpretation of the clinical significance of symptom reduction or lack thereof (Chorpita 2001; Hoagwood et al. 1996; Kazdin and Kendall 1998; Sechrest et al. 1996). By including functioning as a criterion against which to measure our treatments, we can provide more comprehensive information about treatment effectiveness so that clinicians and families select treatments that have the best chance for success in bringing about the outcomes that are most meaningful to them.

Second, mental health care is moving away from a primarily clinician-based model in which a clinician's personal or theoretical preference determines the treatment approach (Dulcan 2005). Advancing in its place is a more industrialized model involving greater attention to quality assurance and client outcomes (Becker et al. 2009; Goldman et al. 2001; Hayes et al. 1999; Schoenwald et al. 2010). With improved quality and accountability standards of managed care organizations and public funding agencies come requirements for assessment of functional impairment to determine service eligibility and measure client outcomes following treatment (Institute of Medicine 2001; Canino et al. 1999; Individuals with Disabilities Education Improvement Act 2004; Plante et al. 1995). Thus, there is the potential for service providers and policy makers to have more leverage in securing reimbursement for evidence-based services that demonstrate consistently positive outcomes (Goldman et al. 2001), particularly on youth functioning.

Third, despite the drive toward quality and client improvement, there continues to be a lack of firm agreement on desired treatment outcomes (Goldman et al. 2001). Managed care and other systems still focus on the cost of services at the expense of resources allocated to delivering services that will bring about improvement in client outcomes that are important to consumers (Hernandez et al. 1998). Psychology as a field has a prime opportunity to influence policy at the state and national levels and help codify client outcomes (Rapp et al. 2005) by expanding our evaluation criteria to include more factors that resonate with mental health stakeholders. This would further influence policy by shaping efforts towards directing funding for the training and implementation of EBTs that demonstrate positive improvements in functioning (Rapp et al. 2005).

Just as our treatments should be subject to evaluation by standards, so too should our standards for evaluation be critically judged (Drake et al. 2003). In this paper, we present our review of 313 RCTs targeting childhood emotional and behavioral problems with the aim of evaluating the effects of applying different criteria for defining evidence-based treatments; specifically, comparisons when outcomes are defined as based on symptom reduction versus improvement in functioning versus their combination. These analyses expand upon a previous paper (i.e., Chorpita et al. in press) that evaluated the strength of evidence for these treatments based solely on the outcome of symptom reduction; thus, the purpose of this paper is to examine the comparative effects of redefining the standard.

Method

The method for identifying treatment outcome studies mirrored the process described in our previous review (i.e., Chorpita et al. in press). Published articles describing RCTs of psychosocial and combined treatments spanning the years from 1965 to 2009 were identified through five methods: (a) computerized searches of electronic databases, (b) review of the reference lists of other major scientific literature reviews (e.g., Silverman and Hinshaw 2008; Weisz et al. 2004), (c) personal communication with national scholars in treatment outcome research, (d) nominations from members of Hawaii's Evidence Based Services Committee (see Chorpita et al. 2002) and the Minnesota Department of Human Services, and (e) email nominations from the professional community to an online, interactive version of earlier reviews. Excluded from our review were universal prevention studies, uncontrolled efficacy trials, studies that described only the follow-up to an RCT, trials that did not test at least one psychosocial treatment, and those studies in which the majority of participants was over the age of 21. Of over 1,500 articles screened for review, 413 articles describing 435 studies had sufficient treatment descriptions, comparative outcome data at post-intervention assessments, evidence of sufficient randomization, and therefore were subjected to detailed coding. Of those coded 435 studies, 373 ultimately met eligibility criteria for analysis (e.g., random assignment, age boundaries), and 313 of those addressed problem areas that were a focus of the current review (references are available upon request). Eight problem areas were included in the current review: anxiety and avoidance, autism spectrum, depression and withdrawal, disruptive behavior, eating problems, inattention and hyperactivity, substance use, and traumatic stress.

Treatments Families

Consistent with our previous reviews (i.e., Chorpita et al. 2002; Chorpita et al. in press), we considered those

interventions that shared a majority of components with similar clinical strategies and theoretical underpinnings as belonging to a single "treatment family" for the purposes of evaluation. For example, rather than score each Cognitive Behavior Therapy protocol for anxiety on its own (we coded over 40 such protocols for this report), these were collectively considered a single group that could achieve a particular level of scientific support. Informed by stakeholder feedback, this approach sought a balance of a reliable separation between constructs of interest in the applied setting and a focus on "generic" as opposed to "brand name" treatment modalities when clear empirical justification for such distinctions did not exist (see Chorpita and Daleiden 2009; Chorpita and Regan 2009). We recognize that proposing definitions of treatment families can introduce subjectivity, and this approach is consistent with recent recommendations from APA Division 53 regarding the importance of identifying and documenting commonalities in efficacious practices that underlie brand name treatments (Rogers and Vismara 2008). Although our coding includes outcome results for waitlist and other control conditions, we restricted the current analyses to "active" treatments, or those interventions intended to have an impact on youth outcomes ($N = 133$ treatment families across all problem areas).

Strength of Evidence

As in our previous review (i.e., Chorpita et al. in press) strength of evidence was defined using a 5-level system, based in part on the original criteria established by the APA Division 12 Task Force for Promotion and Dissemination of Psychological Procedures (1994). Certainly, there exist limitations to any set of standards; thus, this 5-level system is not offered as a definitive view of the literature, but as a collaboratively determined framework reflecting the priorities of stakeholders in children's mental health, including researchers, policymakers, families, and clinicians (see Nakamura et al. 2009). The full set of definitions appears in Table 1.

Indicators of Symptoms and Functioning

Strength of evidence levels were based on treatment outcomes for each study's primary indicator of symptom reduction and functioning. The primary measure was determined by coders according to structured guidelines that prioritized the study authors' designation of the primary measures of symptom reduction and improved functioning. In ambiguous situations regarding which measure was considered "primary," coders selected measures that had

Table 1 Strength of evidence definitions

Level 1: best support
At least two randomized trials demonstrating efficacy in one or more of the following ways:
Superior to pill placebo, psychological placebo, or another treatment
Equivalent to all other groups representing at least one Level 1 or Level 2 treatment in a study with adequate statistical power (30 participants per group on average; cf. Kazdin and Bass 1989) and that showed significant pre-post change in the index group as well as the group(s) being tied. Ties of treatments that have previously qualified only through ties are ineligible
Experiments must be conducted with treatment manuals
Effects must have been demonstrated by at least two different investigator teams
Level 2: good support
Two experiments showing the treatment is (statistically significantly) superior to a waiting-list or no-treatment control group. Manuals, specification of sample, and independent investigators are not required
Or
One between group design experiment with clear specification of group, use of manuals, and demonstrating efficacy by either:
Superior to pill placebo, psychological placebo, or another treatment
Equivalent to an established treatment (see qualifying tie definition above)
Level 3: Moderate support
One between group design experiment with clear specification of group and treatment approach and demonstrating efficacy by either:
Superior to pill placebo, psychological placebo, or another treatment
Equivalent to an already established treatment in experiments with adequate statistical power (30 participants per group on average)
Level 4: minimal support
One experiment showing the treatment is (statistically significantly) superior to a waiting-list or no-treatment control group. Manuals, specification of sample, and independent investigators are not required
Level 5: No support
The treatment has been tested in at least one study, but has failed to meet criteria for level 1 through 4

Note: Reprinted with permission from Chorpita et al. (in press)

greater frequency of occurrence across the literature (a rough indicator of higher standardization and benchmarking ability). When no designation about an instrument's target (i.e., symptoms versus functioning) was provided by the authors, coders selected measures that assessed specific dimensions or indicators of the problem area of interest (for symptoms) or that assessed general (i.e., not specific to a given psychopathological syndrome) impairment in the youth's ability to carry out age-appropriate roles (for functioning). Of note, attendance and grades were coded under a separate "academic" outcome domain, not included in our present definition of functioning.

Outcome indicators for symptom reduction were defined as those that reflected the manifestations of psychopathology (e.g., Winters et al. 2005) that the active treatments in the study were designed to address. Across the 313 studies in this review, 219 different measures of symptoms were used, with the most frequent measures being the Child Behavior Checklist (Achenbach 1991), the Matching Familiar Figures Test (Kagan 1964), and the Revised Children's Manifest Anxiety Scale (Reynolds and Richmond 1985).

Outcome indicators of functioning were defined as those that assessed the youth's ability to meet age-appropriate roles and to participate in life activities without interference (Bird 1999; Hoagwood et al. 1996; Winters et al. 2005). Twenty-six instruments were identified by study authors as indicators of functioning (see Table 2). The Children's Global Assessment Scale (Schaffer et al. 1983) and Vineland Adaptive Behavior Scales (Sparrow et al. 1984) were the two instruments most frequently used to measure functioning. More than three-quarters (76.9%) of the measures used were unidimensional indicators of functioning.

It is recognized that there may be overlap of symptoms and functioning within a single instrument (e.g., DSM-IV criteria for many diagnoses require the presence of specific diagnostic symptoms as well as functional impairment), and there is concern about the validity of existing tools to independently assess one dimension versus the other (Bacon et al. 2002; Goldman 2005). With regard to this controversy, our purpose here was not to add to that debate, but rather to work from the current state of the art regarding how treatments are evaluated along various lines. There exists a body of literature reviewing the validity of assessment tools to measure functioning (e.g., Bacon et al. 2002; Bates 2001; Canino et al. 1999; Winters et al. 2005) and for some instruments, these reviews provide only moderate evidence of validity. Within the present study, the frequency with which these measures were characterized by treatment investigators as indicators of functioning in the studies themselves suggests that there is at least some consensus about what they measure. Despite reasonable concerns about validity, these measures of functioning represent the current standard for measuring the ability of treatments to affect a youth's capacity to meet life expectations without general impairment. We therefore felt that taking such measures at face value was consistent with our goal to investigate the field's ability to establish the potential "value" of treatments, at a given level of measurement precision in the field.

Wins/Ties

Wins/ties represents the actual number of study groups in which a protocol from a particular treatment family performed better than one or more other study groups (a psychosocial treatment, medication, combined psychosocial and medication, placebo, waitlist, no-treatment, or other control group) or tied an evidence-based treatment in an RCT on the primary outcome measure of either the target symptom or functioning domain.

Coding and Reliability

Two raters coded all studies using the PracticeWise Clinical Coding System (PracticeWise 2005) which summarizes multiple variables related to study design, treatment group characteristics, treatment protocols, and results. Each coder received extensive training in the coding system and referred to a detailed coding manual (approximately 70 pp) while coding studies. Coding discrepancies were resolved by an expert reviewer (i.e., second and third authors). Coding and data integrity procedures were identical to those used in our other reviews (see Chorpita and Daleiden 2009 or Chorpita et al. in press for details). Reliabilities for the coded variables relevant to the present study (e.g., use of manual, pre and post *n*, means, and *SD*s), were good and have been reported elsewhere (Chorpita and Daleiden 2009; Chorpita et al. in press).

Results

The results below are organized according to problem area. Within each problem area, the level of evidence is reported for symptom reduction and improvement in youth functioning. First, the total number of identified treatment families is reported, along with the percentage of treatment families that achieved at least level 4 (minimal) support based on symptom reduction (more detailed information regarding strength of evidence levels for symptom reduction may be found in Chorpita et al. in press). Following that, information regarding the number of treatment families that were tested in studies that measured functioning is presented, and for these, the strength of the evidence for

Table 2 Instruments identified by study authors as indicators of functioning

Instrument	Content	Dimensionality	# Studies
Children's Global Assessment Scale	Global functioning	U	14
Vineland Adaptive Behavior Scales			8
Adaptive behavior composite	Daily living skills (e.g., communication, dressing, hygiene)	M	7
Communication scale	Receptive, expressive, and written communication skills	U	1
Global Assessment of Functioning	Global functioning	U	5
Clinical Global Impression Scale-Improvement	Global functioning	U	4
Parent rating	Global functioning	U	2
Professional rating	Global functioning	U	2
Child and Adolescent Functional Assessment Scale			3
Full	Role performance at home, school community; behavior towards others; moods/emotions; self-harmful behavior; substance use; thinking	M	1
Role performance	Role performance at home, school, community	M	1
Role performance: community	Role performance in community	U	1
Health of Nations Outcome Scale for Children and Adolescents			3
Global functioning	Global outcome of recovered or improved functioning	U	2
Global impairment	Global outcome of recovered or improved functioning	U	1
Anxiety Disorders Interview Sched. for DSM-IV-Parent Version			2
Parent rating	Global functioning	U	1
Clinician rating	Global functioning	U	1
Clinician ratings of functioning	Global functioning	U	2
Morgan–Russell Assessment Schedule	Global functioning	U	2
Brief Symptom Inventory: Global Severity Index	Global functioning and distress	U	1
Child Behavior Checklist: Social Competency Scale	Social competency	U	1
Child self-report	Days in locked setting in past 12 months	U	1
Community Adjustment Rating Scale	Overall adjustment	U	1
Global Clinical Rating Scale	Global outcome of recovered or improved functioning	U	1
Mental Health Summary Score 12	Mental health-related quality of life	U	1
Normative Adaptive Behavior Checklist	Total functioning across multiple areas, including fine and gross motor skills, language, home responsibilities, self-help skills	M	1
Pediatric Symptom Checklist	Psychosocial dysfunction as indicated by frequency of emotional and behavioral problems	U	1
Short Form-12: Mental Component Summary	Vitality, social functioning, role limitations due to emotional problems, mental health	M	1
Social Adjustment Scale	Social functioning	U	1
Teacher Report Form: Adaptive Functioning Scale	Hard work, behavior, learning, and happiness at school	M	1

U unidimensional, *M* multidimensional

improving youth functioning is reported. Of the treatment families that were considered "evidence-based" with regard to symptom improvement (i.e., demonstrated at least level 4 support for symptom reduction), the percentage that achieved at least level 4 support for improving youth functioning is reported, along with the percentage of these studies that were not tested in studies that measured functioning and thus were assigned level 5 (no support) with regard to their evidence base for improving youth functioning.

Anxiety and Avoidance

Of 23 identified treatment families targeting anxiety and avoidance, 17 (73.9%) demonstrated at least level 4

Table 3 Leveling results across domains for treatment families targeting anxiety and avoidance

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Cognitive Behavioral Therapy (CBT)	66	66 (100.0)	42 (63.6)	1	9 (13.6)	5 (55.6)	1
Exposure	47	47 (100.0)	32 (68.1)	1	2 (4.3)	1 (50.0)	2 (−)
Modeling	16	16 (100.0)	9 (56.3)	1	0 (0.0)		5 (−)
Education	5	5 (100.0)	3 (60.0)	1	0 (0.0)		5 (−)
CBT plus Medication	2	2 (100.0)	2 (100.0)	1	1 (50.0)	1 (100.0)	2 (−)
CBT with Parents	9	9 (100.0)	3 (33.3)	2	4 (44.4)	2 (50.0)	2
CBT for Child and Parent	3	3 (100.0)	1 (33.3)	2	0 (0.0)		5 (−)
Relaxation	3	3 (100.0)	2 (66.7)	2	0 (0.0)		5 (−)
Assertiveness Training	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Family Psychoeducation	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Hypnosis	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Contingency Management	2	2 (100.0)	1 (50.0)	3	0 (0.0)		5 (−)
Group Therapy	1	1 (100.0)	1 (100.0)	3	0 (0.0)		5 (−)
Play Therapy	3	3 (100.0)	1 (33.3)	4	0 (0.0)		5 (−)
Rational Emotive Therapy	3	3 (100.0)	1 (33.3)	4	0 (0.0)		5 (−)
Biofeedback	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Psychodynamic	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Attachment	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
CBT with Parents Only	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Client-Centered Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Eye Movement Desensitization and Reprocessing (EMDR)	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Relationship Counseling	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Teacher Psychoeducation	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

(minimal) support based on symptom reduction as the outcome indicator of interest (see Table 3). Only 4 (17.4%) of the 23 treatment families were tested in studies that measured functioning; however there was empirical support for all four treatments when defining a positive outcome using a measure of functioning. Cognitive Behavioral Therapy (CBT) retained its level 1 (best) support; however, exposure and CBT plus Medication, which both achieved level 1for symptom reduction, dropped down to level 2 (good) support for improving functioning, joining CBT with Parents, which met level 2 support for both symptom reduction and functioning. As can be seen in Table 11, the majority of treatments considered evidence-based at level 4 or higher for symptom reduction did not achieve level 4 or higher for improving youth functioning. All of the treatments assigned level 5 (no support) for improving youth functioning received no support because they were not tested in studies that measured functioning.

Autism Spectrum

Empirical support at level 4 or higher for symptom reduction was achieved by five of six (83.3%) treatment families identified as targeting autism spectrum (see Table 4). Two of the six treatments (33.3%) were tested in studies that included measures of functioning. Although these two treatments (i.e., Intensive Behavioral Treatment, Intensive Communication Training) demonstrated level 1 support for symptom reduction, neither demonstrated any support based on functioning. None of the treatments that achieved at least level 4 support based on symptom reduction achieved at least level 4 support with regard to

Table 4 Leveling results across domains for treatment families targeting autism spectrum

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Intensive Behavioral Treatment	11	9 (81.8)	4 (44.4)	1	4 (36.3)	0 (0.0)	5 (−)
Intensive Communication Training	4	4 (100.0)	3 (75.0)	1	4 (100.0)	0 (0.0)	5 (−)
Cognitive Behavioral Therapy (CBT)	3	3 (100.0)	1 (33.3)	4	0 (0.0)		5 (−)
Parent Management Training (PMT)	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Peer Pairing	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Auditory Integration Training	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

improving youth functioning, and 60% of them were not tested in studies that measured youth functioning (see Table 11).

Depression and Withdrawal

Of 14 identified treatment families for depression and withdrawal, 10 (71.4%) were considered evidence-based at level 4 or higher based on symptom reduction (see Table 5). Nine of the 14 treatments (64.3%) were tested in studies that included measures of functioning, of which 6 (66.7%) achieved at least level 4 support based on improving youth functioning. CBT and CBT plus Medication retained their level 1 support across target symptoms and functioning, and interpersonal therapy maintained level 2 support. CBT with Parents and Family Therapy dropped from level 1 support for symptom reduction to level 2 support for improving functioning, and Client-Centered

Table 5 Leveling results across domains for treatment families targeting depression and withdrawal

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Cognitive Behavioral Therapy (CBT)	19	19 (100.0)	15 (79.0)	1	7 (36.8)	5 (71.4)	1
CBT plus Medication	5	5 (100.0)	3 (60.0)	1	4 (80.0)	2 (50.0)	1
CBT with Parents	3	3 (100.0)	3 (100.0)	1	2 (66.7)	1 (50.0)	2 (−)
Family Therapy	3	3 (100.0)	2 (66.7)	1	2 (66.7)	1 (50.0)	2 (−)
Interpersonal Therapy	3	3 (100.0)	3 (100.0)	2	2 (66.7)	1 (50.0)	2
Relaxation	3	3 (100.0)	2 (66.7)	2	1 (33.3)	0 (0.0)	5 (−)
Client-Centered Therapy	2	2 (100.0)	1 (50.0)	2	1 (50.0)	1 (100.0)	3 (−)
Expressive Writing/Journaling/Diary	2	2 (100.0)	2 (100.0)	2	0 (0.0)		5 (−)
Self Control Training	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Self Modeling	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Life Skills	1	1 (100.0)	0 (0.0)	5	1 (100.0)	0 (0.0)	5 (−)
Problem Solving	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5 (−)
Psychodynamic Therapy	1	1 (100.0)	0 (0.0)	5	1 (100.0)	0 (0.0)	5 (−)
Social Skills	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5 (−)

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

Therapy dropped from level 2 support for symptom reduction to level 3 support based on improving youth functioning. Three-fifths of the treatments that achieved at least level 4 support based on symptom reduction achieved at least level 4 support based on improving youth functioning (see Table 11). For 75% of the remaining treatments that received no support for improving functioning, lack of support was a result of not being tested in studies that included measures of functioning.

Disruptive Behavior

Empirical support at level 4 or higher for symptom reduction was achieved by 22 of 38 (57.9%) treatment families identified as targeting disruptive behavior (see Table 6). Eleven of 38 treatment families (28.9%) were tested in studies that measured functioning, 8 of which (72.7%) achieved level 4 or higher for improving youth functioning. There were no studies that demonstrated level 1 support for improving youth functioning. Level 2 support was demonstrated by PMT, MST, CBT, and PMT plus problem solving, all of which had achieved level 1 support for target symptoms. Treatment families that maintained their standing of level 2 support across symptoms reduction and youth functioning included Contingency Management, Anger Control, and Therapeutic Foster Care. Collaborative Problem Solving, which received no support for target symptoms, demonstrated level 2 support for functioning. Nearly one-third of the treatments considered evidence-based at level 4 or higher achieved at least level 4 support for improving youth functioning, and the majority of those treatments that did not receive any support were not tested in studies that measured functioning (see Table 11).

Eating Problems

Of 6 treatment families identified as targeting eating problems, 3 (50.0%) demonstrated at least level 4 support for symptom reduction (see Table 7). Four treatment families (66.7%) were tested in studies that included measures of functioning; however, no support for functioning was demonstrated by any of these treatment families. All treatments considered evidence-based with regard to symptom reduction were tested in studies that included indicators of functioning, yet none achieved a win/tie and therefore were not considered evidence-based for improving youth functioning (Table 11).

Inattention and Hyperactivity

Of 25 treatment families identified as targeting inattention and hyperactivity, 16 (64.0%) demonstrated at least level 4 support based on symptom reduction (see Table 8). Only 3 of the 25 treatments (12.0%) were tested in studies that measured functioning, and two out of three (66.7%) demonstrated empirical support for improving youth functioning. Specifically, Parent Management Training (PMT) and Contingency Management both maintained their level 2 support, whereas PMT plus Teacher Psychoeducation received no support (level 5) based on functioning. The majority of treatments considered evidence-based at level 4 or higher for symptom reduction were not considered evidence-based with regard to improving youth functioning (see Table 11). Most of those that failed to achieve level 4 or higher for functioning were not tested in studies that measured functioning.

Substance Use

Eight of 13 treatment families (61.5%) targeting substance use achieved at least level 4 support with regard to symptom reduction (see Table 9). No treatments were tested in studies that included an outcome indicator of youth functioning; thus, none of those treatments that were considered evidence-based with regard to symptom reduction were considered evidence-based with regard to improving youth functioning (see Table 11).

Traumatic Stress

Of eight treatment families identified as targeting traumatic stress, 4 (50.0%) demonstrated at least level 4 support for symptom reduction (see Table 10). Three of eight treatment families (37.5%) were tested in studies that had a measurement of functioning, and all three (100%) demonstrated level 2 support for improving youth functioning. Of note, CBT plus Medication, which had no support for symptom reduction, demonstrated level 2 support for improving youth functioning. Half of the treatments that demonstrated at least level 4 support for symptom reduction were considered evidence-based for improving youth functioning (see Table 11). Neither of the two remaining treatments had support for improving youth functioning, as a result of not being tested in studies that included indicators of youth functioning.

Levels of Evidence for the Combination of Symptom Reduction and Improvement in Functioning

In these analyses, strength of evidence levels were determined by taking into account all of the studies within each problem area in which *either* symptom reduction *or* functioning was measured, such that evidence of efficacy was considered as a whole. (This is quite different from a much more conservative approach to conceptualizing a "combined" category, which would be to consider only those

Table 6 Leveling results across domains for treatment families targeting disruptive behavior

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Parent Management Training (PMT)	58	58 (100.0)	41 (70.7)	1	5 (8.6)	1 (20.0)	2 (−)
Multisystemic Therapy (MST)	12	12 (100.0)	9 (75.0)	1	2 (16.7)	1 (50.0)	2 (−)
Social Skills	11	11 (100.0)	7 (63.6)	1	0 (0.0)		5 (−)
Cognitive Behavior Therapy (CBT)	7	7 (100.0)	4 (57.1)	1	1 (14.3)	1 (100.0)	2 (−)
Assertiveness Training	4	4 (100.0)	3 (75.0)	1	0 (0.0)		5 (−)
PMT and Problem Solving	4	4 (100.0)	3 (75.0)	1	2 (50.0)	1 (50.0)	2 (−)
Problem Solving	12	11 (91.7)	7 (63.6)	2	1 (8.3)	0 (0.0)	5 (−)
Contingency Management	9	9 (100.0)	5 (55.6)	2	1 (11.1)	1 (100.0)	2
Anger Control	7	7 (100.0)	4 (57.1)	2	1 (14.3)	1 (100.0)	2
Communication Skills	5	5 (100.0)	5 (100.0)	2	0 (0.0)		5 (−)
Relaxation	2	2 (100.0)	2 (100.0)	2	0 (0.0)		5 (−)
Therapeutic Foster Care	2	2 (100.0)	2 (100.0)	2	1 (50.0)	1 (100.0)	2
Functional Family Therapy	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
PMT and Classroom Contingency Management	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Rational Emotive Therapy	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Transactional Analysis	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Outreach Counseling	1	1 (100.0)	1 (100.0)	3	0 (0.0)		5 (−)
Peer Pairing	1	1 (100.0)	1 (100.0)	3	0 (0.0)		5 (−)
Self Control Training	1	1 (100.0)	1 (100.0)	3	0 (0.0)		5 (−)
Physical Exercise	2	1 (50.0)	1 (100.0)	4	0 (0.0)		5 (−)
PMT and Self-Verbalization	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Stress Inoculation	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Client-Centered Therapy	7	7 (100.0)	0 (0.0)	5	0 (0.0)		5
Education	4	4 (100.0)	0 (0.0)	5	0 (0.0)		5
Catharsis	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
Family Empowerment	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
Group Therapy	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
CBT Plus Anger Control	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
CBT Child Plus Parents	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Collaborative Problem Solving	1	1 (100.0)	0 (0.0)	5	1 (100.0)	1 (100.0)	2 (+)
Exposure	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Family Systems Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Life Skills	1	1 (100.0)	0 (0.0)	5	1 (100.0)	0 (0.0)	5
Play Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Psychodynamic Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Self-verbalization	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Skill Development	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Wraparound	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

Table 7 Leveling results across domains for treatment families targeting eating problems

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Family Therapy	9	9 (100.0)	2 (22.2)	2	3 (33.3)	0 (0.0)	5 (−)
Cognitive Behavioral Therapy (CBT)	4	4 (100.0)	1 (25.0)	2	1 (25.0)	0 (0.0)	5 (−)
Family Systems Therapy	3	3 (100.0)	2 (66.7)	2	1 (33.3)	0 (0.0)	5 (−)
Education	2	2 (100.0)	0 (0.0)	5	1 (50.0)	0 (0.0)	5
Client-Centered Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Goal-setting	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

studies in which *both* symptom reduction *and* functioning were measured). Thus, this method maximized the opportunity for any treatment family to demonstrate efficacy.

The majority of strength of evidence levels for the combination of symptom reduction or improved functioning mirrored those obtained for symptom reduction alone. Within the problem areas of autism spectrum, eating problems, and substance use, no treatment families achieved a higher level of evidence for the combination than that achieved by symptom reduction alone.

Nevertheless, 7 treatment families achieved a higher level of evidence for the combination than that achieved by symptom reduction. For anxiety and avoidance, the percentage of CBT with Parents groups that demonstrated a win/tie increased from 33.3% (symptom reduction) to 55.6% (combination), resulting in an improvement from level 2 support based on symptom reduction alone to level 1 support for the combination of outcomes. For inattention/hyperactivity, increases from symptom reduction to the combination in the percentage of PMT (41.7–50.0%) and Contingency Management groups (33.3–50.0%) that demonstrated a win/tie resulted in a change from level 2 to level 1 for both groups for the combination. Psychodynamic Therapy for depression and withdrawal demonstrated an increase in the percentage of wins/ties from 0.0% for symptom reduction to 100.0% for the combination, thereby improving its level from 5 to 2. For disruptive behaviors, increases were noted for Contingency Management (55.6–66.7%) and Collaborative Problem Solving (0.0–100.0%). These additional wins/ties allowed Contingency Management to improve its level of support from 2 to 1 and Collaborative Problem Solving to improve its level of support from 5 to 2 from symptom reduction to the combination. Finally, the percentage of wins/ties demonstrated by CBT plus Medication for the treatment of trauma increased from 0.0% for symptom reduction to 100.0% for the combination, allowing an increase in its level of support from 5 to 2.

Conditional Probabilities

Within each problem area, we examined the likelihood that a treatment family would be considered evidence-based (i.e., achieve at least level 4) in one domain (either symptom reduction or improved functioning) if the treatment family had achieved at least level 4 in the other domain. For these conditional probabilities, the number of groups within each problem area that achieved at least one win/tie for symptoms and functioning served as the numerator. The denominator was the number of groups within each problem area that won/tied for symptom reduction and was tested in studies that measured functioning (for the probability of achieving level 4 or higher for improvement in youth functioning) or the number of groups that won/tied for improvement in functioning and was tested in studies that measured symptoms (for the probability of achieving level 4 or higher for symptom reduction).

Additionally, within each problem area, we examined the probability that a treatment that had achieved level 4 or higher (i.e., qualifying win or tie) on one measure of symptom reduction would have a replicated effect on a second measure of symptom reduction, and likewise for one and two measures of improved functioning.

Conditional probabilities are displayed in Table 12. For anxiety/avoidance, inattention/hyperactivity, and depression/withdrawal, when treatments achieved at least level 4 for improving functioning, they also achieved at least level 4 for symptom reduction. Although not quite as high, the conditional probabilities for treatments targeting disruptive

Table 8 Leveling results across domains for treatment families targeting inattention and hyperactivity

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Self Verbalization	13	13 (100.0)	4 (30.8)	1	0 (0.0)		5 (−)
Behavior Therapy plus Medication	4	4 (100.0)	3 (75.0)	1	0 (0.0)		5 (−)
Parent Management Training (PMT)	12	12 (100.0)	5 (41.7)	2	2 (16.7)	1 (50.0)	2
Contingency Management	6	6 (100.0)	2 (33.3)	2	1 (16.7)	1 (100.0)	2
Education	5	3 (60.0)	1 (33.3)	2	0 (0.0)		5 (−)
Biofeedback	4	4 (100.0)	2 (50.0)	2	0 (0.0)		5 (−)
Social Skills plus Medication	4	4 (100.0)	1 (25.0)	2	0 (0.0)		5 (−)
Physical Exercise	3	3 (100.0)	3 (100.0)	2	0 (0.0)		5 (−)
PMT and Teacher Psycho-education	2	2 (100.0)	2 (100.0)	2	1 (50.0)	0 (0.0)	5 (−)
PMT and Problem Solving	2	2 (100.0)	1 (50.0)	2	0 (0.0)		5 (−)
Working Memory Training	2	2 (100.0)	1 (50.0)	2	0 (0.0)		5 (−)
Relaxation and Physical Exercise	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Social Skills	5	5 (100.0)	1 (20.0)	4	0 (0.0)		5 (−)
PMT and Social Skills	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Relaxation	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Self Verbalization and Contingency Management	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Cognitive Behavioral Therapy (CBT)	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
CBT Plus Anger Control	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
Parent Coping/Stress Management	2	0 (0.0)	n/a	5	0 (0.0)		5
Client-Centered Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
PMT and self-verbalization	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Self-control Training	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Self-verbalization Plus Medication	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
Skill Development	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

behavior and traumatic stress demonstrated at least an 80% chance of achieving level 4 for symptom reduction given that they had achieved level 4 for improving functioning. In contrast, for most problem areas, it was about half as likely that a treatment family would achieve at least level 4 for improvement in youth functioning if it had achieved at least level 4 for symptom reduction than vice versa. In other words, it was much more difficult for treatments to achieve level 4 or higher for youth functioning than it was for treatments to achieve the same level of evidence for symptom reduction. The exception to this was that treatment families targeting traumatic stress had the same probability (0.80) of being considered evidence-based in one domain if they were considered evidence-based in the other domain.

The likelihood of replicating the effects for symptom reduction on a second measure ranged from 0.27 (traumatic stress) to 0.80 (eating problems). In contrast, the probability of replicating the effects of improved functioning were much lower and ranged from 0.00 (inattention/hyperactivity) to 0.20 (traumatic stress).

Discussion

Across all problem areas, symptom reduction was clearly the primary outcome indicator in randomized controlled trials targeting childhood behavioral and emotional problems, such that 99.2% of the treatment families were tested in studies that included at least one outcome measure of

Table 9 Leveling results across domains for treatment families targeting substance use

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Family Therapy	6	6 (100.0)	3 (50.0)	1	0 (0.0)		5 (−)
Cognitive Behavioral Therapy (CBT)	4	4 (100.0)	3 (75.0)	2	0 (0.0)		5 (−)
Motivational Interviewing/Engagement	3	3 (100.0)	2 (66.7)	2	0 (0.0)		5 (−)
Family Systems Therapy	2	2 (100.0)	1 (50.0)	2	0 (0.0)		5 (−)
Contingency Management	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Goal Setting/Monitoring	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Purdue Brief Family Therapy	1	1 (100.0)	1 (100.0)	2	0 (0.0)		5 (−)
Goal Setting	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Group Therapy	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
Client-Centered Therapy	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Education	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Project CARE	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Twelve Step Program	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

Table 10 Leveling results across domains for treatment families targeting traumatic stress

Treatment family	Total groups	Symptoms			Functioning		
		Groups (%)[a]	Wins/ties (%)[b]	Level[c]	Groups (%)[a]	Wins/ties (%)[b]	Level (+, −)[d]
Cognitive Behavioral Therapy (CBT) with Parents	5	5 (100.0)	5 (100.0)	1	2 (40.0)	2 (100.0)	2 (−)
CBT	5	5 (100.0)	4 (80.0)	2	3 (60.0)	2 (66.7)	2
Play Therapy	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Psychodrama	1	1 (100.0)	1 (100.0)	4	0 (0.0)		5 (−)
Client-Centered Therapy	2	2 (100.0)	0 (0.0)	5	0 (0.0)		5
CBT + Medication	1	1 (100.0)	0 (0.0)	5	1 (100.0)	1 (100.0)	2 (+)
CBT with Parents only	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5
Eye Movement Desensitization and Reprocessing (EMDR)	1	1 (100.0)	0 (0.0)	5	0 (0.0)		5

[a] Percentages for "Groups" were calculated as Groups (within each domain)/Total groups

[b] Percentages for "Wins/Ties" were calculated as Wins&Ties (within each domain)/Groups

[c] Level represents strength of evidence for each treatment family with regard to symptom reduction. The levels for symptom reduction were previously reported in Chorpita et al. in press

[d] "+" indicates increased level of evidence from the level achieved for symptom reduction only, "−" indicates decreased level of evidence from the level achieved for symptom reduction only

symptoms (the exception to the rule was parent coping/stress management for inattention and hyperactivity problems). In stark contrast, only 26.3% of the treatment families across all problem areas were tested in studies that reported outcomes for youth functioning. There appeared to be some variation with regard to the measurement of functioning across problem areas. For example, RCTs involving treatments for depression and eating disorders were far more likely to include a measure of youth functioning than studies of other youth problems (64.3, 66.7%, respectively). Autism and trauma measured functioning for at least one-third of the treatment families. In contrast, we were unable to identify any RCTs of for youth substance use that included measures of improvement in functioning.

Table 11 Percentage of treatment families achieving level 4 or higher for symptom reduction that achieved or failed to achieve level 4 or higher for improvement in functioning

Target problem	Level 4 or above	Level 5	
		Failure to Win/tie[a]	Failure to measure[b]
Anxiety and Avoidance	23.6	0.0	100.0
Autism Spectrum	0.0	40.0	60.0
Depression and Withdrawal	60.0	25.0	75.0
Disruptive Behavior	30.4	12.5	87.5
Eating Problems	0.0	100.0	0.0
Inattention and Hyperactivity	13.5	7.1	92.9
Substance Use	0.0	0.0	100.0
Traumatic Stress	50.0	0.0	100.0

Note: Level 5 percentages are percent of treatments in column 1 (Level 4 or Above). [a] Treatment family was tested in at least one study that measured youth functioning, but treatment family did not win/tie against comparison group. [b] Treatment family was not tested in at least one study that measured youth functioning

Table 12 Conditional probabilities of a study group achieving a win or qualifying tie in one domain (symptom reduction or improved functioning), given a win or qualifying tie in the other domain and probabilities of replication of effects on a second measure in outcome domain

Problem area	P^a (Sx reduct. improved Fx)	P^b (Improved Fx Sx reduct.)	P^c (Replic. of Sx Reduct. Sx Reduct.)	P^d (Replic. of improved Fx improved Fx)
Anxiety and Avoidance	1.00	0.56	0.59	0.11
Autism Spectrum	–[e]	0.00	0.70	–[e]
Depression and Withdrawal	1.00	0.58	0.52	0.18
Disruptive Behavior	0.88	0.47	0.63	0.13
Eating Problems	–[e]	0.00	0.80	–[e]
Inattention and Hyperactivity	1.00	0.50	0.43	0.00
Substance Use	–[f]	–[f]	0.31	–[f]
Traumatic Stress	0.80	0.80	0.27	0.20

[a] Probability of study group achieving level 4 on symptom reduction, given level 4 or higher for improved functioning

[b] Probability of study group achieving level 4 on improved functioning, given level 4 or higher for symptom reduction

[c] Probability of replication of symptom reduction on a second measure of symptom reduction, given a win/qualifying tie for symptom reduction on one measure

[d] Probability of replication of improved functioning on a second measure of functioning, given a win or qualifying tie for improved functioning on one measure

[e] No treatment families targeting this problem area achieved level 4 for improvement in youth functioning; therefore, probabilities could not be calculated

[f] No treatment families targeting this problem area were tested in studies that included measures of youth functioning; therefore, probabilities could not be calculated

When improvement in youth functioning was measured, most treatments tended to demonstrate at least the minimal level of support, although in far fewer studies than in which functional improvement was measured. Additionally, there was variation across problem areas. Within anxiety and trauma, 100% of the treatment families that were tested in studies in which functioning was measured demonstrated support. At least two-thirds of treatments for attention problems, depression, and disruptive behavior that were tested in studies in which improvement in functioning was measured were found to be evidence-based at level 4 or above for improved functioning. This was not the case for autism and eating disorders, for which 0% of the treatment families demonstrated support for improvement in youth functioning. Though there were notable exceptions of treatment families improving their level of evidence-based level for improvement in functioning compared with that for symptom reduction (e.g., Collaborative Problem Solving for disruptive behavior; Greene et al. 2004; CBT plus medication for traumatic stress; (Cohen et al. 2007), treatment families tended to demonstrate mostly lower levels of empirical support when evidence for improvement in functioning was considered in isolation as compared to the evidence for symptom reduction.

The combination of evidence for symptom reduction and improvement in functioning showed the potential to improve the level of support for some treatments, although this happened for just seven treatment families. Four of these treatment families demonstrated support based on symptom reduction alone and their level of empirical support increased when symptoms and functioning were examined in combination. Of note, the other three treatment families demonstrated support for the combination of symptom reduction and improvement in functioning yet had no support for their efficacy based on target symptoms alone.

Thus, the conclusions are threefold: (1) improvement in youth functioning is measured less frequently than is symptom reduction, (2) it is difficult to demonstrate levels of evidence for improvement in functioning that are comparable to levels of support for symptom reduction, and (3) it is more difficult to demonstrate even minimal levels of empirical support for improvement in functioning than for symptom reduction. That said, there are important messages embedded within the interpretations of these results that reflect broader issues pertinent to psychotherapy research. The development of standards of evidence serves to identify best practices, yet empirical support is not absolute; rather, it is a matter of degree (Westen et al. 2004). Given that the current paradigm for evaluating EBTs is essentially cumulative (i.e., EBTs can only be added to the list), the issue of how to refine our standards is of critical importance. It is likely that maintaining a focus on a unidimensional indicator of empirical support (i.e., symptom reduction) could push our evaluative framework towards obsolescence.

This situation represents a classic signal detection problem (cf. Swets 1992) in which the filter must be set to maximize the signal (in this case, efficacious treatments) while screening out the noise (treatments that are not efficacious). In order to maintain their utility, our standards of evidence require periodic refinements that adapt the criteria by which are treatments are deemed evidence-based to enhance our ability to distinguish among the treatments that populate our growing lists of EBTs. As mentioned earlier, one possible approach is a system of evaluation that involves penalties for null findings. For example, rather than meeting a minimum number of successful tests, treatments would have to be successful in a minimum percentage of published demonstrations. This option would permit continual adjustment based on the state of the literature and would maintain a smaller number of best treatments at the top of the list. However, this approach may be less than ideal due to (1) file drawer effects that could systematically underestimate the number of null findings (McLeod and Weisz 2004), (2) complex decisions about treatment integrity in any studies that produced null findings, and (3) similarly complicated decisions about the scope of generalizability in the event of null findings (e.g., does a null finding imply that a treatment less effective, or only so for a tested subgroup?).

An alternative or potentially complementary approach is to expand the scope of positive outcomes required for "evidence-based" status. Results from the current study converge with the message of others (e.g., Westen and Bradley 2005) that the strength of the evidence depends on the outcome indicator of interest. Asking the question "evidence-based for what purpose?" therefore appears to be reasonable, and designing a system that enhances our ability to answer that question seems to be a worthy endeavor.

Based on the current state of the literature, it appears that the answer to that question is most likely "symptom reduction." Treatments overwhelmingly demonstrated empirical support for symptom reduction compared to improvement in functioning. The results indicate that improvement in functioning represents a higher standard when applied to the literature as a whole. As the conditional probabilities demonstrate, the higher standard is not only due to the measurement effect or burden of administering a functioning measure, but also due to substantively more difficult thresholds for effects.

Across all problem areas, 56.5% of the treatment groups that achieved level 4 or higher for symptom reduction on one measure yielded replicated effects on a second measure of symptom reduction. Replication of effects across problem areas for improved functioning was demonstrated for 25.7% treatment groups. We cannot determine from our data whether the absence of replicated effects reflects the absence of an effect or the absence of a second measure in that particular domain. Nonetheless, the conditional probabilities for the replication of effects call into question the level of confidence that can be placed in an evaluative framework that determines the extent to which treatments are "evidence-based" based on a single outcome indicator.

It apparent that conjunction of improvement in functioning and symptom reduction does not appear to create a fruitful standard because the results are too similar to those demonstrated when considering symptom reduction in isolation to be a practical improvement, so it may be that the addition of functional improvement as a separate standard of evaluation could be a viable solution, although not a perfect one. There exists variation in measurement and empirical support across problem areas. For example, imposing a standard of functional improvement for treatments targeting anxiety would result in the identification of four EBTs, all of them with a CBT/exposure focus. The same standard applied to eating problems would result in no treatments identified as evidence-based. It is possible that different standards could be applied to different

problem areas, but that would increase the complexity of the evaluative framework. Perhaps a more practical solution might be to add an even higher level to identify those treatments that have empirical support based on symptom reduction that have also passed the additional threshold for evidence of improved functioning. However, this option may be less than ideal because it relegates our standards for evaluation back to essentially a single dimension and does not provide a framework that will easily adapt to an evolving literature.

Another layer of complexity with regard to using symptom reduction and functional improvement together to provide a standard of evidence has to do with the conceptual distinction of these constructs but their intermingling within instruments that purport to measure one or the other. The validity of the measurement of functioning has been discussed in numerous articles over the last 20 years (e.g., Canino et al. 1999; Goldman et al. 1992; Winters et al. 2005). One facet of the debate involves the validity of unidimensional versus multidimensional scales. Global measures of functioning provide a common metric across disorders, providers, and youth (Burlingame et al. 1995), yet clinician ratings on unidimensional scales may be heavily influenced by symptom severity (Bacon et al. 2002). Multidimensional scales may provide useful information about functioning in a specific context, yet certain multidimensional scales have unknown or questionable psychometric properties and tend to include indicators of symptomatology (Bates 2001; Bickman et al. 1998; Winters et al. 2005).

Unfortunately, the lack of strong support for the validity of measures of functioning (e.g., Bacon et al. 2002; Bates 2001; Canino et al. 1999; Winters et al. 2005) undermines our ability to determine the full universe of EBTs that have the intended impact of improving the ability of youth to carry out their daily roles and responsibilities in the face of adversity. At the same time, reliance on measures of functioning with questionable validity to determine levels of evidence may provide a more rigorous standard such that only the best treatments from the best designed studies would be able to demonstrate improvement. A related issue regarding measures of functioning has to do with sensitivity to change (Canino et al. 1999). Some research suggests that improvement on measures of functioning may lag behind improvement on measures of symptoms (Bickman et al. 1999), so it may be that improvement in functioning following a time-limited treatment should also reflect a more rigorous standard of evidence than symptom reduction. Certainly, our ability to make inferences about the potential for treatments to achieve their intended social impact is limited by the measurement precision within the field, an issue that should improve over time as measurement of functioning becomes better understood and operationalized.

Even so, the inclusion of one additional outcome indicator to improve the utility of our evaluative framework may lead to only incremental improvement in evaluative accuracy that will diminish over time. Instead, perhaps the best way to optimize a strength of evidence rating system would be to base judgments of evidence on the empirical support demonstrated by a treatment across multiple dimensions, including symptom reduction, functioning, satisfaction, ecology, etc. (see Hoagwood et al. 1996). A multidimensional framework has the potential to shape the clinical and empirical landscape in many ways. First, the transparency of a multidimensional framework would provide the opportunity for our field to increase the validity of the "evidence-based" designation. Second, it would provide an opportunity to increase the mental health treatment literacy of stakeholders so that they can be informed participants in personal or policy decisions about treatment. Third, the establishment of such a framework would further support clinical decision-making by identifying primary treatment options based on traditional symptom reduction heuristics while allowing for the consideration of other dimensions (e.g., functioning) when there is clinical equipoise. Fourth, a multidimensional standard of evaluation would encourage psychotherapy researchers developing and studying EBTs to expand their outcomes to be more representative of an array of outcomes as well as to harness their scientific ingenuity to enhance existing treatments in ways that improve their outcomes across multiple dimensions rather than symptom reduction alone. In general, a multidimensional framework has the potential to provide a dynamic approach to treatment evaluation with the ability to keep pace with the state of the literature, thereby enhancing its long-term value.

Limitations

Over the last 50 years, treatment pioneers have been developing, testing, and refining treatments to address childhood problems with much success. It is certainly a testament to the efforts in the field that we are at the point at which we are not only able to evaluate treatments, but to re-evaluate our methods of evaluation. Any systematic review of treatment studies naturally is shaped by strengths and limitations of the literature, as well as the difficult decisions that are inherent in the creation of standards for evaluation; therefore, our results must be interpreted within this context. Evident from our review is a literature comprised of intervention studies that reflect increasingly stronger methodologies over time. Our exclusion of studies with designs other than RCT is generally consistent with previous reviews in the field (cf. Chambless et al. 1996, 1998; Chorpita et al. 2002; Task Force on Promotion and Dissemination of Psychological Procedures 1995), but this

approach has been criticized because it omits other methodological approaches such as multiple baseline designs, quasi-experimental designs, and case studies and also because some treatments purportedly may be more amenable to manualization or RCT testing than others (Westen et al. 2004). Although it is not without limitations (see Drake et al. 2004; Westen et al. 2004), the randomized controlled trial is held as the gold standard for interventions research across fields and provides the best opportunity to measure efficacy (Drake et al. 2004; Weisz et al. 2005).

Due to limitations that reflect the current state of the literature, the criteria for each evidence level in our framework may require refinement in the future as intervention studies continue to improve their methodology. For example, our evidence levels were based on comparisons between the active treatment and any control condition (e.g., waitlist, attention, active treatment), although comparisons between two active treatments would provide the most stringent test of efficacy. However, of those studies included in our review, relatively few treatments were tested against other active treatments. If the trend continues such that RCTs are designed to include active comparison treatments, 1 day it may be possible to consider "comparison against an active treatment" as an additional criterion that distinguishes a level 1 from a level 2 treatment.

Similarly, our strength of evidence framework provides a snapshot of treatment efficacy at posttreatment only and cannot provide information regarding the long-term efficacy of any EBTs. Although long-term efficacy may be an important piece of information considered within clinical decision-making, there are too few follow-up studies of treatments to be synthesized using the present approach at this time, but it is certainly the case that this framework could be applied to follow-up studies as that literature grows. Additionally, follow-up studies may enhance the opportunity for treatments to achieve empirical support based on functioning (when such measures are administered in a study) because it may be the case that improvements in functioning lag behind symptom reduction (Bickman et al. 1999).

Our evaluative approach also assumed a single problem area, rather than the comorbid problems that characterize the majority of youth that receive treatment. Again, this reflects the overall state of the literature, specifically with regard to the historical focus on single disorder study samples within RCTs. However, our emphasis on problem area, rather than diagnosis per se, provided a more liberal criterion for RCTs to be coded in our study. Specifically, we sought to include RCTs with samples of youth exhibiting subclinical problems or specified risk factors for specific problem areas. As research samples become increasingly diverse with regard to comorbid presenting problems, our evaluative framework can be adjusted to reflect outcomes that correspond with the nature of the comorbidity.

Finally, our evidence level designations were based on the raw number of wins/ties achieved by each treatment. Wins were based on statistical significance (and lack thereof, for ties). There was no consideration of effect sizes or percent of children improved, which are important indicators of treatment efficacy in their own right (Jacobson et al. 1999). Effect sizes for symptom reduction were presented in our previous review (i.e., Chorpita et al. in press), although evidence levels were not based on effect sizes in that study either. It is possible that effect sizes could inform strength of evidence criteria in future reviews.

Conclusion

Throughout this paper, we have attempted to present our data in a way that highlights not only the importance of evaluating treatments, but also the importance of reassessing the standards by which we evaluate those treatments. Given that the working definitions of evidence were developed when over-identification was not a challenge to the field, we view these issues as representative of steady progress in the field of psychotherapy research. We hope and expect that psychotherapy research will continue to contribute to the ever-increasing array of effective psychosocial treatments, and thus, we feel that the time is upon us to question what standards would look like that could endure and serve the field for the next 20 years. Regardless of their content, it is already a significant triumph that our best treatments may soon outgrow the very standards that our field so recently established to define them.

Acknowledgment During the preparation of this manuscript, Dr. Becker was supported by the National Institute of Mental Health (T32 MH18834).

Disclosure Bruce F. Chorpita is President and Eric L. Daleiden is Chief Operating Officer of PracticeWise, LLC, a private behavioral health consulting corporation. Kimberly D. Becker is a consultant to PracticeWise, LLC.

References

Achenbach, T. (1991). *Manual for the child behavior checklist/4–18 and 1991 profile*. Burlington: Department of Psychiatry, University of Vermont.

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

Bacon, S., Collins, M., & Plake, E. (2002). Does the Global Assessment of Functioning assess functioning? *Journal of Mental Health Counseling, 24*, 202–212.

Bates, M. (2001). The Child and Adolescent Functional Assessment Scale (CAFAS): Review and current status. *Clinical Child and Family Psychology Review, 4*, 63–84.

Becker, K. D., Nakamura, B., Young, J., & Chorpita, B. (2009). What better place than here? What better time than now? ABCT's burgeoning role in the dissemination and implementation of evidence-based practices. *The Behavior Therapist, 32*, 89–96.

Bickman, L., Lambert, E., Karver, M., & Andrade, A. (1998). Two low-cost measures of child and adolescent functioning for services research. *Evaluation and Program Planning, 21*, 263–275.

Bickman, L., Lambert, E., & Schilling, S. (1999). Time patterns of mental health outcomes. *Mental Health Services Research, 1*, 67–82.

Bickman, L., Rosof-Williams, J., Salzer, M., Summerfelt, W., Noser, K., Wilson, S., et al. (2000). What information do clinicians value for monitoring adolescent client progress and outcomes? *Professional Psychology: Research and Practice, 31*, 70–74.

Bird, H. (1999). The assessment of functional impairment. In D. Shaffer, C. Lucas, & J. Richters (Eds.), *Diagnostic assessment in child and adolescent psychopathology* (pp. 209–229). NY: Guilford.

Bird, H., Yager, T., Staghezza, B., & Gould, M. (1990). Impairment in the epidemiological measurement of childhood psychopathology in the community. *Journal of the American Academy of Child & Adolescent Psychiatry, 29*, 796–803.

Burlingame, G., Lambert, M., Reisinger, C., Neff, W., & Mosier, J. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *Mental Health Services Research, 22*, 226–236.

Canino, G., Costello, E., & Angold, A. (1999). Assessing functional impairment for child mental health services research: A review of measures. *Journal of Mental Health Services Research, 1*, 93–108.

Chambless, D., Baker, M., Baucom, D., Beurtler, L., Calhoun, K., Crits-Cristoph, P., et al. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist, 51*, 3–16.

Chambless, D., Sanderson, W., Shoham, V., Bennett Johnson, S., Pope, K., Crits-Christoph, P., et al. (1996). An update on empirically validated therapies. *The Clinical Psychologist, 49*, 5–18.

Chorpita, B. (2001). Reflections on clinical significance: What do our best treatments accomplish and how can we best find out? *Clinical Psychology: Science and Practice, 8*, 451–454.

Chorpita, B., & Daleiden, E. (2009). Mapping evidence-based treatments for children and adolescents: Application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology, 77*, 566–579.

Chorpita, B., Daleiden, E., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. et al. (in press). Evidence based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical Psychology: Science and Practice*.

Chorpita, B., & Regan, J. (2009). Dissemination of effective mental health treatment procedures: Maximizing the return on a significant investment. *Behaviour Research and Therapy, 47*, 990–993.

Chorpita, B., Yim, L., Donkervoet, J., Arensdorf, A., Amundsen, M., McGee, C., et al. (2002). Toward large-scale implementation of empirically supported treatments for children: A review and observations by the Hawaii Empirical Basis to Services Task Force. *Clinical Psychology: Science and Practice, 9*, 165–190.

Cohen, J., Mannarino, A., Perel, J., & Staron, V. (2007). A pilot randomized controlled trial of combined trauma-focused CBT and sertraline for childhood PTSD symptoms. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*, 811–819.

De Los Reyes, A., & Kazdin, A. (2008). When the evidence says, 'yes, no, and maybe so': Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science, 17*, 47–51.

Drake, R., Latimer, E., Leff, H., McHugo, G., & Burns, B. (2004). What is evidence? *Child and Adolescent Psychiatric Clinics of North America, 13*, 717–728.

Drake, R., Rosenberg, S., Teague, G., Bartels, S., & Torrey, W. (2003). Fundamental principles of evidence-based medicine applied to mental health care. *Psychiatric Clinics of North America, 26*, 811–820.

Dulcan, M. (2005). Practitioner perspectives on evidence-based practice. *Child and Adolescent Psychiatric Clinics of North America, 14*, 225–240.

Goldman, H. (2005). Do you walk to school, or do you carry your lunch? *Psychiatric Services, 56*, 419.

Goldman, H., Ganju, V., Drake, R., Gorman, P., Hogan, M., Hyde, P., et al. (2001). Policy implications for implementing evidence-based practices. *Psychiatric Services, 52*, 1591–1597.

Goldman, H., Skodol, A., & Lave, T. (1992). Revising axis V for DSM-IV: A review of measures of social functioning. *American Journal of Psychiatry, 149*, 1148–1156.

Greene, R., Ablon, J., Monuteaux, M., Goring, J., Henin, A., Raezer, L., et al. (2004). Effectiveness of collaborative problem solving in affectively dysregulated youth with oppositional defiant disorder: Initial findings. *Journal of Consulting and Clinical Psychology, 72*, 1157–1164.

Hayes, S., Barlow, D., & Nelson-Gray, R. (1999). *The scientist practitioner: Research and accountability in the age of managed care. Needham heights* (2nd ed.). MA US: Allyn & Bacon.

Hernandez, M., Hodges, S., & Cascardi, M. (1998). The ecology of outcomes: System accountability in children's mental health. *The Journal of Behavioral Health Services & Research, 25*, 136–150.

Hoagwood, K., Jensen, P., Petti, T., & Burns, B. (1996). Outcomes of mental health care for children and adolescents: I. A comprehensive conceptual model. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*, 1055–1063.

Hodges, K., Doucette-Gates, A., & Kim, C. (2000). Predicting service utilization with the child and adolescent functional assessment scale in a sample of youths with serious emotional disturbance served by center for mental health Services-funded demonstrations. *The Journal of Behavioral Health Services & Research, 27*, 47–59.

Individuals with Disabilities Education Improvement Act (2004) H.R. 1350, 108th Congress.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Jacobson, N., Roberts, L., Berns, S., & McGlinchey, J. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67*, 300–307.

Kagan, J. (1964). *Matching familiar figures test*. Cambridge, MA: Harvard University.

Karver, M., & Bickman, L. (2002). Positive functioning: Does it add validity to maladaptive functioning items? *Evaluation and Program Planning, 25*, 85–93.

Kazdin, A., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138–147.

Kazdin, A., & Kendall, P. (1998). Current progress and future plans for developing effective treatments: Comments and perspectives. *Journal of Clinical Child Psychology, 27*, 217–226.

McGlinchey, J., Zimmerman, M., & Atkins, D. (2008). Clinical significance and remission in treating major depressive disorder: Parallels between related outcome constructs. *Harvard Review of Psychiatry, 16*, 25–34.

McLeod, B., & Weisz, J. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. *Journal of Consulting and Clinical Psychology, 72*, 235–251.

Nakamura, B., Chorpita, B., Hisch, M., Daleiden, E., Slavin, L., Amundson, M. et al. (2009). *Large-scale implementation of evidence-based treatments for children ten years later: Hawaii's evidence based services initiative in children's mental health* (submitted).

Pickles, A., Rowe, R., Simonoff, E., Foley, D., Rutter, M., & Silberg, J. (2001). Child psychiatric symptoms and psychosocial impairment: Relationship and prognostic significance. *British Journal of Psychiatry, 179*, 230–235.

Plante, T., Couchman, C., & Diaz, A. (1995). Measuring treatment outcome and client satisfaction among children and families. *Journal of Mental Health Administration, 22*, 61–269.

PracticeWise. (2005). Psychosocial and combined treatments coding manual. Satellite Beach, FL: Author.

Rapp, C., Bond, G., Becker, D., Carpinello, S., Nikkel, R., & Gintoli, G. (2005). The role of state mental health authorities in promoting improved client outcomes through evidence-based practice. *Community Mental Health Journal, 41*, 347–363.

Reynolds, C., & Richmond, B. (1985). *Revised children's manifest anxiety scale*. Los Angeles: Western Psychological Services.

Rogers, S., & Vismara, L. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child & Adolescent Psychology, 37*, 8–38.

Schaffer, D., et al. (1983). A children's global assessment scale (CGAS). *Archives of General Psychiatry, 40*, 1228–1231.

Schoenwald, S., Hoagwood, K., Atkins, M., Evans, S., & Ringeisen, H. (2010). Workforce development and the organization of work: The science we need. *Administration and Policy In Mental Health, 37*, 71–80.

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist, 51*, 1065–1071.

Silverman, W., & Hinshaw, S. (2008). The second special issue on evidence-based psychosocial treatments for children and adolescents: A 10-year update. *Journal of Clinical Child and Adolescent Psychology, 37*, 1–7.

Simonoff, E., et al. (1997). The Virginia twin study of adolescent behavioral development: Influences of age, gender and impairment on rates of disorder. *Archives of General Psychiatry, 54*, 801–808.

Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland adaptive behavior scales: Interview edition survey form manual*. Circle Pines, MN: American Guidance Service.

Striley, C., Stiffman, A., & Spitznagel, E. (2003). Functioning mediates between symptoms and provider assessment. *Mental Health Services Research, 5*, 155–171.

Swets, J. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532.

Task Force on Promotion, Dissemination of Psychological Procedures D. o. C. P. (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist, 48*, 3–23.

Wampold, B., Imel, Z., & Miller, S. (2009). Barriers to the dissemination of empirically supported treatments: Matching messages to the evidence. *The Behavior Therapist, 32*, 144–155.

Weisz, J., Hawley, K., & Jensen-Doss, A. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent Psychiatric Clinics of North America, 13*, 729–815.

Weisz, J., Weersing, V., & Henngeler, S. (2005). Jousting with straw men: Comment on Westen, Novotny, and Thompson-Brenner. *Psychological Bulletin, 131*, 418–426.

Westen, D., & Bradley, R. (2005). Empirically supported complexity: Rethinking evidence-based practice in psychotherapy. *Current Directions in Psychological Science, 14*, 266–271.

Westen, D., Novotny, C., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin, 130*, 631–663.

Winters, N., Collett, B., & Myers, K. (2005). Ten-year review of rating scales, VII: Scales assessing functional impairment. *Journal of the Academy of Child and Adolescent Psychiatry, 44*, 309–338.

Zimmerman, M., McGlinchey, J. B., Posternak, M. A., Friedman, M., Attiullah, M., & Boerescu, D. (2006). How should remission from depression be defined? The depressed patient's perspective. *The American Journal of Psychiatry, 163*, 148–150.