Annu. Rev. Psychol. 2005. 56:337–63
doi: 10.1146/annurev.psych.55.090902.141449

First published online as a Review in Advance on September 10, 2004

YOUTH PSYCHOTHERAPY OUTCOME RESEARCH: A Review and Critique of the Evidence Base

John R. Weisz
Judge Baker Children's Center, Harvard University, Boston, Massachusetts 02120-3225; email: jweisz@jbcc.harvard.edu

Amanda Jensen Doss
Department of Educational Psychology, Texas A & M University, College Station, Texas 77843-4225

Kristin M. Hawley
Department of Psychological Sciences, University of Missouri, Columbia, Missouri 65211

Key Words children, adolescents, psychotherapy, mental health, treatment

■ **Abstract** Over the past four decades, researchers have produced extensive evidence on psychotherapy for youth mental health problems and disorders. The evidence often has been evaluated through narrative reviews and through meta-analyses assessing the magnitude of treatment effects, but methodological analysis addressing the character and quality of the evidence base itself is an important complement, needed to place treatment effects in perspective and to suggest directions for future research. We carried out such an analysis, focusing on all the methodologically acceptable published randomized trials our search identified involving treatment of anxiety, depression, ADHD and related conditions, and conduct-related problems and disorders. The 236 studies tested 383 treatments and included 427 treatment-control comparisons, spanning the years 1962 through 2002. The analysis revealed considerable breadth, diversity, and rigor in the measurement approaches used to assess participant characteristics and treatment outcomes. However, reporting on important sample characteristics (e.g., ethnicity) showed major gaps, and more than half the studies failed to use well-standardized procedures to ensure appropriate sample selection. Because sample sizes left most studies underpowered, and procedures to enhance treatment fidelity were generally weak, many of the treatments investigated may not have received fair tests. Studies were particularly weak in clinical representativeness of their samples, therapists, and settings, suggesting a need for increased emphasis on external validity in youth treatment research.

CONTENTS

INTRODUCTION: BRIEF HISTORY OF YOUTH PSYCHOTHERAPY AND OUTCOME RESEARCH

The term "psychotherapy" refers to an array of nonmedical interventions designed to alleviate nonnormative psychological distress, reduce maladaptive behavior, or increase deficient adaptive behavior through counseling, interaction, a training program, or a predetermined treatment plan. One likely source of this practice is the ancient tradition of helping by listening, discussing, and questioning (see Plato's *Apology*). An early practitioner, Socrates, developed both a method and a thesis that arguably set a pattern for some modern forms of psychotherapy. His approach, later called the Socratic method, involved questioning others to prompt examination of their beliefs and bring them closer to truth. His "midwife thesis," the notion that the philosopher's role was to deliver the truth that is already within others, much like the midwife delivers the baby that is within the mother, is not far from the view many modern therapists have of their own professional roles. By asking others to tell him what they thought, rather than telling them what to think, Socrates sought to reach the rational soul or psyche of those he talked with. The term psyche denoted the mind, inner nature, and capacity for feeling, desire, and reasoning, and was the precursor to the word psychology. Finally, Socrates maintained that thoughts and outward behavior are closely connected, presaging a tenet of many modern therapies.

Just when psychotherapy became a career track is not clear, but a case can be made for the era when Sigmund Freud (1856–1939) launched psychoanalysis. Important themes in this work were the notions that early experience can be critical, and that even children may be appropriate candidates for intervention. These points were illustrated by Freud's (1909) treatment of a boy ("Little Hans") who was afraid of horses by consulting with the boy's father. Freud's daughter Anna (1895–1982) later became a prominent child analyst, as did many others in the first half of the twentieth century. Other models and methods propelled the acceleration of child psychotherapy through the century, including a radically different behavioral approach. Mary Cover Jones (1924), for example, used modeling and gradual exposure to help a two-year-old, Peter, overcome fear of a white rabbit. This work helped to launch a remarkable burgeoning of behavioral psychotherapies for young people, complementing psychoanalysis and humanistic treatments. By the late twentieth century, child and adolescent psychotherapy had expanded remarkably in the variety of its forms and the extent of its reach.

The growth of psychotherapy prompted increasing curiosity about its impact. Although psychotherapy research developed later and more slowly than psychotherapy practice, studies began to accumulate. Eysenck (1952) reviewed studies of adult psychotherapy and concluded that the evidence did not show it to be effective. A few years later, Levitt (1957, 1963) reviewed studies that included children and adolescents and concluded that rates of improvement in the youth were about the same with or without treatment. These early reviews were influential, but the studies they relied on were not rigorous by today's methodological standards. Subsequent research has grown stronger and much more plentiful (Durlak et al. 1995, Kazdin 2000). Many of the studies now in the evidence base meet the standards of randomized clinical trials, and their focus has sharpened over the decades, shifting from early studies of unspecified "treatment" for often vaguely defined youth problems to tests of rather well-articulated therapies targeting specific patterns of dysfunction. In sum, thanks to several historical developments, we are now in a position to profit from a large and increasingly rigorous body of evidence on youth psychotherapies and their effects.

PREVIOUS NARRATIVE AND QUANTITATIVE REVIEWS

The evidence on youth psychotherapies is examined periodically through both narrative and quantitative reviews of outcome findings. Narrative reviews (e.g., Kazdin 2000, Shirk & Russell 1996) can bring the perspectives of thoughtful experts to bear on what the outcome findings show, and can identify the strengths and limitations of those findings. As an example, Kazdin's (2000) scrutiny of the research on youth treatment led to a detailed critique of what has and has not been learned, and a list of specific recommendations for improving the yield of youth outcome trials. As another example, Shirk & Russell (1996) focused their narrative review partly on the question of what treatment research has told us about

pathogenic processes and change processes in treatment. This review, too, led to a proposed new framework for individualizing and implementing youth treatments.

Another approach to reviewing outcome research is the quantitative review of treatment effects, an approach called meta-analysis. Meta-analysts apply a common effect size metric to sets of studies, to gauge the magnitude of treatment benefit (or harm) for entire bodies of evidence or selected subsets. Four published meta-analyses of the youth treatment outcome literature have been particularly broad in their inclusion criteria, encompassing studies of diverse problems and disorders and varied treatments (Casey & Berman 1985; Kazdin et al. 1990; Weisz et al. 1987, 1995). Across these four meta-analyses, mean unweighted effect sizes were all 0.71 or higher, indicating that the average treated youth scored better than more than three fourths of control group youths on outcome measures at the end of treatment. Effect sizes are more modest, but still positive and substantial, when weighting is introduced to adjust for sample size (Weisz et al. 1995). Meta-analytic findings also suggest that treatment effects show specificity—i.e., effects are larger on measures of the problems actually targeted in treatment than on measures of other mental health outcomes not targeted—and that the effects hold up well after the end of treatment, at least for the five- to six-month follow-up period that is common in youth outcome research (Weisz et al. 1995).

QUANTITATIVE METHODOLOGICAL ANALYSIS

A third method of taking stock of the field is quantitative methodological analysis. In this approach, the focus shifts to the nature and quality of the evidence base from which narrative and quantitative reviewers derive their findings and conclusions. One goal is to lend perspective to those findings and conclusions. As an example, conclusions about the potency of treatment programs in comparison to control groups are best evaluated in light of information about what those treatment programs were, how faithfully they were carried out, to whom they were delivered, and with what strength, as well as information about the kinds of control groups to which they were compared. Summary statements about outcome findings are also best framed by an understanding of the kinds of outcome measures used, who provided the information gathered via those measures, and the extent to which the measures and those providing them were apt to be free of bias. More broadly, examining the evidence base may provide insights into the lay of our land in youth treatment outcome research, and whether investigators are gathering the kinds of information most needed to understand and improve mental health care for young people.

In one example of quantitative methodological analysis, Kazdin et al. (1990) reviewed published studies of psychotherapy for youngsters aged 4–18. In another example, Durlak et al. (1995) reviewed psychotherapy studies with children aged 13 and younger. By examining features of their respective study sets, Kazdin et al. and Durlak et al. were able to identify important characteristics of the evidence base. They noted, for example, that the research disproportionately involved tests

of behavioral and cognitive-behavioral procedures, with significant underrepresentation of other treatment approaches that are widely practiced. In addition, the reviewers noted that outcome studies in general focused primarily on testing techniques, with scant attention to nontechnique variables (such as youth or family characteristics) that might have a substantial impact on treatment outcome. Through their observations about the structure and content of the treatment outcome studies in their collection, these investigators characterized the evidence base upon which so many conclusions of so many reviewers had rested, and in the process influenced the direction of future youth outcome research.

Because Kazdin and colleagues (1990) reviewed studies published between 1970 and 1988, and the Durlak et al. (1995) review encompassed studies from more than a decade ago, the content of the evidence base is now likely to have changed significantly. In this chapter, we report the results of our own methodological analysis, focused on studies published from 1963 through 2002. In addition to our expanded time range, we encompass the adolescent years not included by Durlak et al., and we introduce new inclusion requirements not previously employed (e.g., requiring random assignment to treatment and control groups) to ensure the methodological rigor of studies. Another distinguishing feature of our approach is that we focus attention on four particularly important clusters of youth dysfunction, as discussed in the following section.

Focusing on Four Common Problem/Disorder Clusters

Youth psychotherapy is used to address diverse problems and disorders that cause emotional distress, interfere with daily living, undermine the development of adaptive skills, or threaten the well-being of others. The concerns addressed may include, for example, enuresis, Tourette's syndrome, anorexia nervosa, fire setting, and trichotillomania. Encompassing every treated condition within a quantitative methodological review could pose a risk that conclusions lack clear referents within the treatment outcome literature. To reduce this risk, we chose to focus on four broad problem/disorder clusters that appear to account for a very high percentage of youth referrals for mental health care in the United States (see, e.g., Jensen & Weisz 2002). The four clusters are:

- Anxiety-related problems and disorders (e.g., social phobia, generalized anxiety disorder)
- Depression-related problems and disorders (e.g., dysthymic disorder, major depressive disorder)
- Attentional problems, impulsivity, and attention deficit/hyperactivity disorder (ADHD)
- Conduct-related problems and disorders (e.g., oppositional defiant disorder, conduct disorder)

With a focus on these four clusters, we examined the treatment outcome database in an effort to address several issues of significance for the field.

Characteristics of Study Samples

We examined both the size and the demographic characteristics of outcome study samples. One aim was to determine whether the sample sizes generally employed have generated sufficient power to detect treatment effects. Another aim was to assess the range of human characteristics tapped in the studies and thus ascertain whether certain groups of young people are missing. We also hoped to learn about the quality of reporting on participant characteristics and the extent to which researchers have been conscientious about clearly characterizing important dimensions of human variability (e.g., ethnicity) in their samples. Attention to such dimensions is relevant to our ability to identify moderators of treatment outcome and to judge how representative the clinical trial samples are of those populations seen in clinical service settings.

Assessment of Diagnoses, Target Problems, and Intervention Outcomes

We examined the process by which investigators identified youths as appropriate for their treatment programs. In the process, we hoped to learn how much confidence to place in the fit of youth treatments to the problems addressed by those treatments. To shed light on the degree of rigor and objectivity used in assessing treatment effects, we also noted the measurement approaches used, focusing on both the informants who provided outcome information and the measurement technology employed.

Another focus in our examination of outcome assessment was breadth of coverage. Several investigators (e.g., Hoagwood et al. 1996, Kazdin 2000, Weisz 2004) have argued that youth treatment research needs to examine intervention impact in ways that go beyond assessment of the specifically targeted youth symptoms and diagnoses. To assess the extent to which this has happened in research to date, we examined how much attention has been given to nontarget youth outcomes (e.g., reduced depressive symptoms in treatment for ADHD), real-world functioning by the youths (e.g., school performance), "environmental impact" (e.g., reduced depressive symptoms in parents of children treated for conduct problems), and client satisfaction with treatment.

Treatments Tested and Control Conditions to Which They Have Been Compared

We also examined the array of treatments tested. We sought to learn the extent to which they focused on youths themselves versus others in the youths' environments who might play key roles in family, school, or community. In addition, we noted the theoretical orientations represented by the tested treatments. We have been concerned over the years (e.g., Weisz et al. 1987, Weisz & Hawley 1998) that the treatment models tested most often in research are not the models most widely used in practice, and that this results in a very limited picture of how well the most

widely used interventions work. We generated data on the extent to which this problem is evident in a particularly comprehensive collection of methodologically sound studies focused on four of the most clinically significant problem clusters.

In addition, we took steps to characterize the treatments with respect to their strength or "dose," and the formats most commonly employed. This characterization included the extent to which youths, parents, entire families, and teachers have been included in treatment sessions, and the steps taken to ensure skill building by youths (i.e., the extent to which therapeutic homework has been assigned) and to ensure the faithful delivery of treatment programs by therapists. All these elements speak to the likely potency of the treatments as delivered and tested. Because the potency measurement of a treatment depends in part on the control condition to which it is compared, we also examined the kinds of control groups employed in treatment trials.

Clinical Representativeness

Finally, a long-standing concern of ours (Weisz & Hawley 1998, Weisz et al. 2004) has been that a significant proportion of treatment outcome research may involve youths, therapists, and settings that are rather different from those of everyday clinical practice, and that this difference may limit the relevance of the research to practice. The evidence base available at the time this review was conducted provided an opportunity to assess the validity of these concerns for research on four of the most common forms of youth dysfunction targeted in everyday clinical care.

METHODS EMPLOYED IN THE REVIEW

To provide the most reliable and relevant evidence bearing on the questions noted above, we sought to identify treatment outcome studies meeting uniform standards that were important for our particular purposes; this made our study collection different from those used in previous reviews. Here we describe the search process, the inclusion criteria we applied, and the coding system used to characterize the studies.

Literature Search Procedures

To identify relevant studies, we used a variety of sources. First, we searched standard computerized databases beginning in 1965 and eventually continuing through the end of 2002. We used PsycInfo, employing 21 psychotherapy-related key terms (e.g., psychother-, counseling, treatment) derived from previous youth psychotherapy meta-analyses (Weisz et al. 1987, 1995), and we used MEDLINE via PubMed, the principal bibliographic database of the National Library of Medicine. PubMed uses a controlled vocabulary indexing system (MeSH) that provides a consistent way to retrieve citations from publishers who may use different key words for the same concepts. We used "mental disorders," with the following search limits: clinical trial, child (3–18 years), published in English, and human subjects.

The MeSH descriptor, "mental disorders," encompasses all related subject terms more specific than this broad heading (e.g., anxiety, mood disorders, mental disorders, ADHD). In addition to these database searches, we surveyed published reviews and meta-analyses of the youth psychotherapy and pharmacotherapy literature to identify studies not found in PsycInfo or MEDLINE (e.g., Casey & Berman 1985, Compton et al. 2002, Durlak et al. 1991, Dush et al. 1989, Farmer et al. 2002, Prout & DeMartino 1986). We also followed reference trails of reviewed studies, and we screened studies suggested by investigators in the field.

We obtained copies of all published youth treatment outcome studies identified by the procedures noted above. To ensure some level of quality control, we selected only studies that had been subjected to peer review. In most cases, this meant that the studies had been published in peer-reviewed journals; we excluded unpublished manuscripts, book chapters (except in the rare cases where serious peer review was evident), and unpublished dissertations. Applying this criterion meant that our ultimate collection of studies would not fully represent all research that has been carried out (see McLeod & Weisz 2004). However, our goal was to characterize the outcome research that met our minimum criteria for methodological soundness (e.g., random assignment) and that had been judged acceptable in professional peer review and made available to the field through publication. We included all studies that met inclusion criteria regardless of treatment outcome.

To be included in the review, studies were required to be tests of psychotherapy, defined as any intervention designed to alleviate nonnormative psychological distress or reduce maladaptive behavior through counseling, interaction, a training program, or a predetermined treatment plan. Studies were also required to:

- Include comparison of psychotherapy to a control group (waitlist, no treatment, placebo, or other process intended not to be an active treatment)
- Involve prospective design and random assignment of subjects to treatment and control conditions
- Use a sample within the 3- to 18-year range
- Use participants selected for having psychological problems or maladaptive behavior (within the four problem clusters noted above)
- Include a posttreatment assessment of the psychological problem(s) or maladaptive behavior for which participants were selected and treated

To ensure a focus on comparison of psychotherapy to a control group, we selected for review only those studies in which participants in the groups to be compared were not taking psychotropic medications.

Coding Procedures and Intercoder Reliability

After we identified studies, we coded characteristics of their samples, settings, treatments, treatment providers, design characteristics, and types of outcomes assessed. To establish interrater reliability for these codes, project coders (one

postdoctoral fellow who served as the master coder and two clinical psychology graduate students) coded 30 randomly selected studies. To compute reliability, each graduate student coder's results were compared to those of the master coder, with kappa (k) statistics computed for categorical codes and Pearson correlation coefficients (r) computed for continuous codes. The codes used and the corresponding intercoder reliabilities are detailed below.

PARTICIPANT/SAMPLE/LOCATION INFORMATION We coded study samples for (a) size of treatment and control groups (mean $r = 0.99$), (b) age at treatment onset (mean $r = 0.99$), (c) gender (percentage of sample that was male: mean $r = 0.99$), (d) ethnicity (percentage of sample that was Caucasian, African American, Latino Hispanic, Asian Pacific, other; mean $r = 0.93$), (e) recruitment source (recruited, clinically referred, court-referred; mean $k = 0.71$). We also coded for the location where treatment was administered (western, midwestern, southern, or eastern United States, Caribbean/U.S. territory, international; mean $k = 0.87$).

TARGET PROBLEM INFORMATION We classified the problems targeted in treatment into four types: (a) anxiety, fears, and shyness; (b) depression; (c) impulsivity and ADHD-related problems; and (d) oppositional-conduct problems (mean $k = 0.91$).

PROBLEM ASSESSMENT INFORMATION We coded the investigators' procedures for assessing and identifying participants' problems. The coding included diagnosis given for target problem (whether participants met diagnostic criteria, via the prevailing diagnostic system, for a disorder falling within the target problem domain; mean $k = 0.79$) and clinical cutoff met for target problem (i.e., whether participants met a clinical cutoff on a standardized continuous measure, such as a self-report symptom questionnaire; mean $k = 0.56$).

TREATMENT AND CONTROL METHODS USED Each randomly assigned treatment or control group was coded according to treatment or control technique. Coders classified study treatment groups with reference to the primary participant or target of the intervention (e.g., child, family, parent, teacher) and the theoretical orientation (e.g., operant, respondent- or exposure-based, cognitive-behavioral, psychodynamic, client-centered, systems-based, eclectic, other). Coders classified control groups into (a) no therapy/waitlist, psychotherapy placebo (active condition designed to control for nonspecific effects), (b) medication placebo (sugar pill or equivalent), and (c) case monitoring or management (e.g., standard, custodial, educational, or judicial). Interrater reliability for the classification of treatment and control conditions resulted in a mean k of 0.75 (treatment and control group designations were assigned categorically as a single code).

FORMAT, NATURE, AND INTENSITY OF TREATMENT Treatment conditions were coded for format by noting (a) whether sessions were individual or group (mean $k = 0.80$), (b) whether or not clients received homework (mean $k = 0.67$), and

(*c*) which form(s) of treatment contact were involved (i.e., contact between target youth and therapist, youth's parent(s) and therapist, youth's family and therapist, youth's teacher and therapist) (mean $k = 0.60$). In addition, treatment dose was coded in terms of (*a*) total number of sessions (mean $r = 0.99$), (*b*) total weeks of treatment (mean $r = 0.98$), (*c*) average length of sessions (mean $r = 0.99$), and (*d*) total hours in treatment (mean $r = 1.00$).

EFFORTS TO ENSURE TREATMENT INTEGRITY We coded (*a*) whether investigators arranged for pretherapy training in the specific therapy techniques that would be used (mean $k = 0.94$), (*b*) whether adherence checks were used (mean $k = 0.74$), and (*c*) whether a treatment manual or equivalent documentation was employed (mean $k = 0.71$).

TREATMENT SETTING The setting where treatment took place was coded to indicate if it was (*a*) a typical clinical service setting (e.g., community clinic versus a university lab office), (*b*) a typical research setting (e.g., university lab office, school), or (*c*) a correctional setting (e.g., prison); some settings were difficult to classify because of incomplete reporting in the article (mean $k = 0.53$).

THERAPIST PRIMARY VOCATION We coded whether treatments were provided by therapists of the following primary vocations: practicing mental health clinician, graduate student or researcher, or paraprofessional (neither a researcher nor a mental health clinician, e.g., teacher); therapist vocation was sometimes difficult to classify because of incomplete reporting in the article (mean $k = 0.63$).

OUTCOME MEASURE CHARACTERISTICS We coded outcome measures with regard to the domain the measure assessed, including symptoms or diagnosis, functioning, environmental impact of treatment (e.g., parent self-esteem or mood), and satisfaction with treatment (mean $k = 0.83$). If the measure was in the domain of symptoms or diagnosis, we further coded whether it assessed the target problem for which the sample had been selected and treated (mean $k = 0.89$).

METHODOLOGICAL CHARACTER AND QUALITY OF THE MEASURES To gauge the methodological character and strength of the measures employed, we coded each measure in each study for (*a*) measurement technology (self-report, other report, objective behavior count, independent life event data) (mean $k = 0.87$), and (*b*) source or reporter (target child, parent/guardian, sibling, peer, teacher, therapist, other observer), (mean $k = 0.89$). We also coded for whether the source or reporter was aware of the target youth's treatment status at the time the outcome measure was obtained (i.e., blindness of source = yes/no) (mean $k = 0.67$). In addition, we coded for whether the subject of the rating (generally the child) was aware that a measure was being taken and was reasonably able to influence the

results of that rating (blindness of subject = yes/no; e.g., observer coding of child behavior during a five-minute assessment task would be coded as yes, assuming that the child was aware of the assessment; observer coding of naturally occurring child behavior every day at recess over two weeks would be coded as no) (mean $k = 0.76$).

RESULTS: DESCRIBING THE EVIDENCE BASE

Although our initial search led to a pool of more than 3000 studies focused on treatment, the methodology of many of the studies was not strong enough to support credible inferences about treatment effects. Application of our methodological criteria (e.g., random assignment required) reduced the pool to 236 methodologically acceptable studies, which had tested 383 different treatments and included 427 treatment-control comparisons. This total included 82 studies in the anxiety domain, testing 137 anxiety treatments and employing 162 treatment-control comparisons; 18 studies in the depression domain, testing 27 depression treatments and employing 28 treatment-control comparisons; 40 studies in the ADHD domain, testing 74 different treatments and employing 84 treatment-control comparisons; and 96 studies in the conduct problem domain, testing 145 different treatments for conduct problems and employing 153 treatment-control comparisons.

Across the full set of studies, year of publication spanned four decades, from 1962 to 2002, with a median publication year of 1986; 39% of the studies were published in the 1990s or later. The anxiety studies were published from 1967 to 2002; median year was 1985; 43% were published in the 1990s or later. The depression studies were published from 1986 to 2002; median was year 1997; 89% were published in the 1990s or later. ADHD studies spanned the years 1968 to 2001; median was year 1981; only 20% were published in the 1990s or later. Finally, for conduct studies, publication ranged from 1963 to 2001; median year 1986; 33% published in the 1990s or later.

Study Samples: Demographic and Characteristics

Table 1 summarizes demographic characteristics of the study samples, with data presented for the four problem clusters and the total collection of studies. The table shows that mean sample age was lowest for ADHD and highest for depression studies, perhaps reflecting early the detection of ADHD as children enter educational settings and the fact that base rates of depression are low in childhood but increase with the transition to adolescence. Samples were predominantly male for the two externalizing clusters, and predominantly female for anxiety and depression. Studies in which race and ethnicity were reported indicated that Caucasian youths were the majority sampled in all four problem clusters, particularly in the case of ADHD. Substantial percentages of black youths were sampled across the four clusters as well, but there was much more limited representation of other

TABLE 1 Demographic characteristics of treatment study samples*

	Anxiety (N = 82)	Depression (N = 18)	ADHD (N = 40)	Conduct (N = 96)	All studies (N = 236)
Mean age in years	10.17	13.87	8.53	10.53	10.30
Studies not reporting	1.22	0	0	2.08	1.27
Boys in study	44.31	40.76	85.63	75.65	64.03
Studies not reporting	20.73	0	17.50	11.46	14.83
Ethnicity					
Caucasian	56.14	57.48	82.80	50.88	56.56
Black	24.38	19.98	12.59	33.04	26.88
Latino	6.20	9.63	0.00	4.66	5.18
Asian	2.88	3.16	0.00	0.14	1.25
Other	7.97	6.22	1.61	10.96	8.56
Studies not reporting	68.29	33.33	75.00	51.04	59.75
Income/SES					
Studies reporting income	0	0	2.50	11.46	5.08
Reporting SES	0	0	2.50	6.25	2.97
Studies not reporting	82.93	88.89	77.50	61.46	72.46
Geographic region					
Studies in United States	58.52	61.11	75.00	73.97	67.80
International studies	36.59	16.67	12.50	18.75	23.73
Studies not reporting	4.88	22.22	12.50	7.29	8.47

*All values (except mean age) are noted in percentages.

ADHD, attention deficit/hyperactivity disorder; SES, socioeconomic status.

ethnic groups. Although many in the field are greatly interested in the income and socioeconomic status (SES) of the samples, we are not able to provide any useful substantive summary because reporting was absent in most cases and erratic across studies that did provide such information (with some reports of income, some of SES, others of the percentage of family members receiving public assistance, etc.). The data on geographic region make it clear that the youth treatment evidence base is largely an American product, with more than two thirds of all the studies conducted in the United States. The most "international" of the four domains is anxiety, for which more than one third of the evidence comes from other countries—particularly Australia, Canada, Israel, the Netherlands, and the United Kingdom. In general, a striking feature of Table 1 is the fact that so much information needed to characterize and understand the youths and families targeted in treatment is simply absent from the written reports. As an example, 60% of all the articles failed to include any report on the race or ethnicity of their samples, and more than 70% failed to provide any information on family income or socioeconomic status.

Procedures Used to Identify Youth Problems and to Assess Outcomes

Table 2 summarizes the procedures used to identify youths as having the diagnoses or problems that were to be targeted in studies. Formal Diagnostic and Statistical Manual (DSM) or International Classification of Diseases (ICD) diagnoses were used to select some or all of the sample in approximately one fourth of all studies; diagnoses were most often used in depression studies, least often in conduct studies. However, even within the minority of studies that did use DSM diagnoses, fewer than half used standardized diagnostic procedures, raising questions about the validity of the diagnoses for which some of the youngsters were selected and treated. An alternate, nondiagnostic approach to problem identification is the use of clinical cutoff scores on standardized continuous measures of psychopathology. Although a strong case can be made for the value of this sample selection strategy, it has been used even less than categorical diagnosis. More than 80% of studies failed to demonstrate that even some of their sample met inclusion criteria based

TABLE 2 Measurement: I. Identification of youth problems and kinds of measures obtained*

	Anxiety	Depression	ADHD	Conduct	All studies
Mean number target problems in measures/study	5.39	4.56	5.60	4.62	5.05
Target problem confirmed via DSM diagnosis?					
Studies w/DSM for full sample	25.61	50.00	32.50	8.33	21.61
Studies w/DSM for partial sample	2.44	0	0	6.25	3.39
Studies w/DSM for none of sample	71.95	50.00	67.50	85.42	75.00
Among studies using DSM, total using standardized procedures	82.60	22.22	38.50	21.40	49.15
Target problem confirmed via clinical cutoff on standardized continuous measure?					
Studies w/cutoff for full sample	6.10	22.22	30.00	14.58	14.83
Studies w/cutoff for partial sample	1.22	11.11	2.50	2.08	2.54
Studies w/cutoff for none of sample	92.68	66.67	67.50	83.33	82.63
Studies not employing DSM or cutoff	68.29	27.78	50.00	73.96	64.41
Mean number of nontarget problem in measures/study	3.26	8.83	4.90	10.16	6.77
Studies w/nontarget symptom measures	75.61	94.44	55.00	86.46	77.97
Studies w/functioning measures	32.93	11.11	22.50	30.21	28.39
Studies w/environment impact measures	1.22	5.56	2.50	8.33	4.66
Studies w/client satisfaction measures	4.88	11.11	2.50	11.46	7.63

*All values (except mean number of target and nontarget problems) are noted in percentages.

ADHD, attention deficit/hyperactivity disorder; DSM, Diagnostic & Statistical Manual.

on a clinical cutoff. Combining both sets of codes revealed that more than 60% of all studies failed to employ either approach for even a portion of the sample. Given the low rates of sample selection via formal diagnosis and via clinical cutoffs, how did most study authors identify youths for participation in their clinical trials? In fact, most studies identified participants based on teacher or parent reports on nonstandard measures or questionnaires, or via responses to an advertisement or to researcher requests for referrals of youths with particular kinds of problems; when DSM diagnoses were used, they were most often assigned by clinicians or researchers using clinical judgment and other nonstandardized procedures of unknown validity. The one domain in which sample selection was carried out more carefully and systematically was depression; 72% of depression studies used either standardized DSM diagnostic procedures or clinical cutoffs on standardized measures. This may reflect the fact that the depression literature includes more recent studies than the other areas (as discussed previously), thus sample selection profited from the more rigorous methodological standards of the era.

Also provided in Table 2 is information about the number and types of measures used in the evidence base. In contrast to the rather limited use of formal diagnosis and even clinical cutoffs to identify study samples, the evidence shows extensive and diverse measurement of participants. Across the four problem domains, studies have averaged approximately five different measures of target problems and approximately seven different measures of nontarget problems; thus, the average study has obtained twelve participant measures. By our inclusion criteria for the review, 100% of the studies included at least one measure of the target problem. In addition, though, 78% of all studies included at least one measure of nontarget problems or symptoms (e.g., a measure of depressed mood or self-esteem obtained in a study of ADHD treatment), 28% of all studies included at least one functioning measure (e.g., school grades or disciplinary incidents), 8% included at least one consumer satisfaction measure, and 5% included at least one measure of the environmental impact of treatment, broadly construed (e.g., maternal depression or parenting stress assessed in a study testing behavioral parent training for child conduct problems). Clearly, the range in types and content of measurement is broad across studies. Equally clearly, though, questions such as how would the participants evaluate the treatment and what impact did treatment have on everyday functioning in school or at home have not been of primary interest to most treatment researchers over the years.

Informants and Technology Employed in Assessment

Table 3 lists the informants used, and the measurement technology employed, in assessing participants and their treatment outcome. It is evident from the table that the youths themselves are highly favored as informants when anxiety and depression are the treatment targets, but are much less likely to be informants in studies focused on ADHD and conduct problems. Because treatment of ADHD and conduct problems focuses on outwardly observable behavior, the treated youths' own

TABLE 3 Measurement: II. Sources of information and measurement technology*

	Anxiety	Depression	ADHD	Conduct	All studies
Sources of information in measures used					
Studies using treated youth as reporter	76.83	94.44	37.50	47.92	59.75
Studies using parent as reporter	28.05	55.56	25.00	35.42	32.63
Studies using sibling as reporter	0	0	0	0	0
Studies using peers as reporters	1.22	0.00	2.50	4.17	2.54
Studies using teachers as reporters	13.41	5.56	37.50	42.71	28.81
Studies using therapist as reporter	9.76	5.56	5.00	6.25	7.63
Studies using other reporters/observers	71.95	77.78	87.50	79.17	77.97
Measurement technology					
Studies using self-report	75.61	94.44	35.00	64.58	65.68
Studies using other report	43.90	83.33	50.00	80.21	62.71
Studies using behavior counts	69.51	38.89	90.00	51.04	63.14
Of these, total using observer ratings	85.97	85.70	94.44	95.92	76.56
Of these, total using blind raters	46.94	33.33	38.24	46.81	44.12
Of these, total in which youth was not aware of or able to influence rating	30.61	100.00	20.59	78.72	47.79
Studies using independent life event data	8.54	5.56	2.50	26.04	14.41

*All values are noted in percentages.

reports may be less critical to assessment than in cases of anxiety and depression, for which treatment focuses on internal processes that are hidden from view. Parents often have been tapped for their perspective, particularly in depression studies, whereas teachers were used as reporters most often in studies focused on the two externalizing clusters. Table 3 also shows heavy reliance on "other reporters," which are generally clinical assessors and trained observers. Interestingly, in none of the 236 studies did we find any reliance on what might be a particularly well-informed source of information on treated youths: their siblings.

A second aspect of Table 3 is its summary of the kinds of measurement technology employed in youth treatment research. Use of informants' self-reports and collateral reports by others, typically accomplished via questionnaires and interviews, has been common. In addition, to an encouraging extent, there has been substantial use of what appear to be more objective approaches, including behavior counts and behavioral observation (e.g., heart rate, performance on a skill test, observer recordings of out-of-seat behavior and prosocial behavior), and independent life event data (e.g., arrests, school attendance records). Enthusiasm about this measurement approach is tempered somewhat by other data in the table: Of those studies that included behavioral observation, fewer than half reported the use of raters who were blind to the treatment condition of the youths they observed; this was the case across all studies and within each of the four problem clusters.

Moreover, with the exception of depression studies, a substantial proportion of the studies using behavioral observation did so with youths who knew they were being observed and who could have intentionally influenced the ratings.

Types of Treatment Tested in the Evidence Base

Table 4 summarizes the kinds of treatments that have been tested to date. It is clear that youth-focused treatments have been much more common than parent- and family-focused treatments, even for the externalizing problems and disorders for which parent training programs are so well known. The table also makes it clear that behavioral or "learning-based" treatments are included in 8 to 10 times as many studies as insight-based approaches, broadly construed. The fact that insight-based methods are so poorly represented in the treatment of anxiety is especially interesting in light of the important historical role such treatments have had in relation to anxiety (e.g., Freud 1909). Of course, this historical emphasis was on the use of such methods, not on their scientific study. By contrast, the development of behavioral interventions was closely linked to the experimental tradition, and that linkage is reflected, no doubt, in the extensive representation of behavioral methods within the youth treatment outcome research literature, as shown in Table 4.

Treatment Characteristics: Dose, Format, Participants, Homework, Integrity

In addition to the theoretical models and therapeutic procedures associated with tested treatments, it is useful to examine treatment strength or "dose," the format and participants involved, and other aspects related to the intensity and care with which interventions were carried out. Table 5 provides information relevant to these issues. The upper portion of the table shows that the average treatment involved 11 to 12 one-hour sessions, spanning 8 to 9 weeks, and totaling about 13 hours of contact. The lengthiest treatments have tended to be those addressing conduct problems and depression. Data on treatment formats show that individual treatment has been more common than group treatment for ADHD, but that the reverse has been true for depression and conduct problems. Youths themselves are by far the most frequent participants in treatment (included within treatment sessions in 83% of studies), but parents often have been involved (25%), as have entire families (18%), and, more rarely, teachers (8%).

Table 5 also outlines the steps taken to build mastery by treated individuals and to support faithful delivery of treatments by therapists. Client homework assignments have been a part of treatment in a majority of studies focused on anxiety (65%), depression (83%), and conduct problems (59%), but less so in ADHD treatment trials (30%); only 40% of the full study set reported any use of homework. Data on the steps taken to foster treatment integrity revealed mixed support. Only 52% of studies reported the use of a treatment manual, but an additional 33% described treatments that we rated as "structured"—e.g., outlining the content covered in

TABLE 4 Theoretical orientation and type of treatment provided*

	Anxiety	Depression	ADHD	Conduct	All studies
Child-focused treatments					
Studies w/learning-based approaches	90.24	77.78	90.00	45.83	71.19
Studies w/operant treatment	3.66	0	7.50	9.38	6.36
Studies w/respondent treatment	25.61	11.11	15.00	2.08	13.14
Studies w/modeling treatment	13.41	5.56	15.00	0	7.63
Studies w/CBT treatment	43.90	66.67	52.50	26.04	39.83
Studies w/social skills treatment	4.88	11.11	2.50	8.33	6.36
Studies w/multiple or other approaches	13.41	0	17.50	5.21	9.75
Studies w/insight-based approaches	7.32	11.11	2.50	10.42	8.05
Studies w/client-centered treatment	7.32	0	2.50	4.17	4.66
Studies w/psychodynamic treatment	0	0	0	4.17	1.69
Studies w/gestalt treatment	0	0	0	0	0
Studies w/multiple or other approaches	1.22	11.11	0	2.18	2.12
Studies w/eclectic approaches	1.22	5.56	0	9.38	4.66
Parent-focused treatments					
Studies w/learning-based approaches	1.22	0	7.50	18.75	9.32
Studies w/insight-based approaches	0	0	0	0	0
Studies w/eclectic approaches	0	0	0	2.08	0.85
Family-focused treatments					
Studies w/learning-based approaches	2.44	5.56	0	10.42	5.51
Studies w/insight-based approaches	0	5.56	0	2.08	1.27
Studies w/systems-based approaches	0	0	0	1.04	0.42
Studies w/eclectic approaches	0	0	0	0	0
Teacher-focused treatments					
Studies w/eclectic approaches	0	0	0	1.04	0.42
Multiple target/multisystem treatments					
Studies w/child + parent	4.88	11.11	10.00	4.17	5.93
Studies w/child + family	3.66	5.56	2.50	0	2.12
Studies w/child + parent + family	3.66	5.56	0	1.04	2.12
Studies w/child + teacher	0	0	0	1.04	0.42
Studies w/child + parent + teacher	0	0	0	1.04	0.42
Studies w/parent + teacher	0	0	2.50	1.04	0.85
Studies w/child + parent + family + teacher	0	0	0	5.21	2.12

*All values are noted in percentages.

ADHD, attention deficit/hyperactivity disorder; CBT, cognitive behavior therapy.

TABLE 5 Treatment characteristics: dose, format, participants, homework, integrity*

	Anxiety	Depression	ADHD	Conduct	All studies
Treatment dose					
Mean number of sessions	9.61	12.26	9.24	14.38	11.44
Mean number of weeks	6.69	9.47	5.79	11.53	8.58
Mean length of sessions in minutes	53.01	67.53	42.15	64.98	57.35
Mean total hours of treatment	9.96	14.33	8.59	17.61	12.92
Format of treatment sessions					
Studies using individual sessions	45.12	33.33	62.50	31.25	41.53
Studies using group sessions	45.12	61.11	22.50	52.08	45.34
Studies using individual + group	6.10	11.11	7.50	7.29	7.20
Studies not reporting format	6.10	16.67	12.50	18.75	13.14
Treatment participants					
Studies involving youths in sessions	97.56	88.89	90.00	66.67	83.05
Studies involving parents in sessions	17.07	27.78	17.50	33.33	24.58
Studies involving families in sessions	14.63	27.78	7.50	23.96	18.22
Studies involving teachers in sessions	4.88	5.56	10.00	10.42	8.05
Studies w/any homework assigned	64.63	83.33	30.00	59.38	40.25
Steps to support treatment integrity					
Studies w/pretreatment therapist training	26.83	66.67	22.50	34.38	32.20
Studies w/supervision/adherence checks	24.39	66.67	32.50	32.29	32.20
Studies using treatment manuals	47.56	72.22	52.50	52.08	52.12
Studies reporting structured treatments	35.37	16.67	45.00	29.17	33.05

*All values (except those listed under Treatment dose) are noted in percentages.

ADHD, attention deficit/hyperactivity disorder.

sessions. Only 32% of the studies noted any pretreatment training for therapists, and only 32% noted any use of supervision procedures or adherence checks. This is another area where depression studies as a group surpassed studies in the other three areas; 67% of depression studies employed pretreatment training, 67% used supervision procedures or adherence checks, and 89% used either treatment manuals or structured treatments. It is possible that treatment adherence support was higher than we could detect, if some authors simply failed to note these aspects of their procedure. Nevertheless, the evidence presented in the articles we reviewed raises concerns about the level of fidelity of actual treatments to the protocols tested. This, in turn, suggests the possibility that some of the interventions delivered may have been weaker than, or at least different from, what the treatment developers and researchers intended.

Treatment Groups and Control Groups

Table 6 shows that mean sample size across all trials in our review was 22 for treatment groups and 21 for control groups. The highest means, 30 and 31

TABLE 6 Treatment and control groups*

	Anxiety	Depression	ADHD	Conduct	All studies
Mean sample size of treatment groups	18.23	30.41	12.38	26.31	21.95
Mean sample size of control groups	16.78	31.41	11.66	24.36	20.62
Types of control groups					
Studies using no treatment/waitlist	64.63	77.78	42.50	64.58	61.86
Studies using attention/placebo	39.02	27.78	70.00	29.17	39.41
Studies using medication placebo	0	0	0	0	0
Studies using standard case management	4.88	0	0	14.58	7.63

*All values (except mean sample sizes) are noted in percentages.

respectively, for treatment and control groups in depression trials, still fall well below the standard of 50 cases per treatment/control group noted by Chambless & Hollon (1998; based on Cohen 1988). Table 6 also shows that the most common type of control group employed in youth trials is waitlist or no treatment. The one exception to this rule is the research base on ADHD treatment. For ADHD studies alone, control groups involving attention and/or placebo interventions were actually more common than no treatment or waitlist control groups (70% versus 43%).

Clinical Representativeness of the Evidence Base

Table 7 shows the extent to which the treatment research is congruent with clinical practice along three different dimensions. In general, levels of clinical representativeness in the evidence base are low. About 13% of the study samples were actually treatment-seeking or clinically referred youth; about 19% of the studies employed at least one practicing clinician (although for many of these studies, practitioners were a minority of the therapists used); only about 4% of the studies provided treatment within a clinical service setting. Clinical representativeness was highest, across all three dimensions, for studies treating depression and conduct problems. Anxiety studies were least representative of clinical practice conditions. Perhaps the most complete picture of representativeness is presented by the figures at the bottom of the table, which show that only about 1% of the total study set was rated as representative across all three dimensions.

OVERVIEW AND CRITIQUE OF THE EVIDENCE BASE, AND SUGGESTIONS FOR IMPROVING FUTURE RESEARCH

This methodological review of the evidence base has highlighted several trends, and some important limitations, in youth outcome research, with respect to four broad problem clusters.

TABLE 7 Clinical representativeness of the studies: youths, therapists, and settings*

	Anxiety	Depression	ADHD	Conduct	All studies
How youths were enrolled in the study					
Recruited, not treatment-seeking	90.24	77.78	87.50	60.42	76.69
Treatment-seeking, clinic-referred	3.66	16.67	12.50	19.79	12.71
Required via court/justice system	1.22	0	0	17.71	7.63
Studies not reporting	4.88	5.56	0	2.08	2.97
Who provided the treatment					
With any researchers/grads	57.32	47.06	45.00	38.54	47.21
With any paraprofessionals	20.73	11.11	12.50	22.92	19.49
With any practicing clinicians	1.22	55.56	10.00	30.21	18.64
Studies not reporting	28.05	11.11	40.00	19.79	25.42
Setting where treatment took place					
Research settings	50.00	44.44	42.50	48.96	47.88
Clinical service settings	2.44	5.56	0	7.29	4.24
Correctional settings	1.22	0	0	7.29	3.39
Studies not reporting	46.34	50.00	55.00	37.50	44.49
Representativeness sum (youths, therapists, and settings)					
Reporting no representativeness factors	92.68	38.89	77.50	55.21	70.76
Reporting one representativeness factor	7.32	50.00	22.50	34.38	24.15
Reporting two representativeness factors	0	5.56	0	8.33	3.81
Reporting all three representativeness factors	0	5.56	0	2.08	1.27

*All values are noted in percentages.

HISTORICAL ERAS AND THE INFLUENCE OF THEORIES From an historical perspective, a kind of era effect is evident. The year of publication data we presented showed that controlled research on problems within the anxiety, ADHD, and conduct clusters borders on the ancient, with randomized trials in each domain found as early as the mid 1960s. By contrast, the first published randomized trial for youth depression appeared two decades later, in 1986, and the median year of publication for depression trials was 1997. This trend may suggest something about the power of theories to dictate targets of study. As Hammen & Rudolph (1996) note, early psychoanalytic theory held that depression in children was an impossibility because the superego was not sufficiently well developed to direct aggression against the self (see also Rochlin 1959). Even in the late 1970s, writers not associated with psychoanalytic theory also argued against the existence of depression in children (Lefkowitz & Burton 1978). In the early 1980s, views shifted, and evidence was assembled that appeared to support the existence of a depression syndrome in children and adolescents (see Cantwell & Carlson 1983). With the relatively recent "discovery" of depression in young people, treatment research has begun, but it may be decades before there is a substantial body of work. Interestingly, however,

one effect of the delayed start in the depression domain appears to be that the methodology used in depression studies as a whole is superior to that in other domains in several important respects, as we noted above.

This review also illustrates the fact that high levels of interest, theory, and literature devoted to a particular theoretical perspective on dysfunction do not automatically translate into high levels of intervention research. Consider anxiety, for example. It has the oldest history of theoretical and clinical attention of any of the youth problems (see, e.g., Freud 1909), with a great deal of the early work reflecting psychoanalytic thinking and other psychodynamic models. Yet, we came up empty in our search for randomized clinical trials of psychodynamic treatments for anxiety in youth. Clearly, theory and research do not invariably stimulate one another. On the other hand, when theory is closely linked to research, or when research is embedded in the value system of a theory, the picture may be very different. This may help to explain the overwhelming number of randomized trials testing learning-based treatments, in contrast to the much lower numbers for other theoretical orientations.

ASSESSMENT AND DOCUMENTATION OF SAMPLE CHARACTERISTICS Although this review yielded a substantive picture of treatment research, one particularly striking finding was the very spotty sample description across studies. Indeed, some potentially important information that would have been readily available at the time studies were done is now lost, with no opportunity for recovery. As an example, race and ethnicity were not reported for the samples in about 60% of the studies. Gaps of this sort can hamper efforts to understand how well treatments work across a range of human characteristics. Experts in minority mental health have expressed serious concerns about whether evidence from ethnic minority groups has in fact supported evidence-based treatments. The minority supplement to the U.S. Surgeon General's report on mental health (U.S. Department of Health and Human Services 2001) noted that the ethnicity of a substantial percentage of participants was not identified in treatment outcome studies. Our figures on the youth treatment research literature support this concern: 60% of the published studies in our review failed to provide ethnicity data. Chambless et al. (1996), reporting on the work of a task force on empirically supported treatments, commented that they did not know of any psychotherapy treatment research that met basic criteria for demonstrating treatment efficacy for minority populations. Our review suggests that this strong conclusion may not (any longer) hold for youth treatment research. The percentages of studies in our pool that did report sample ethnicity point to substantial representation of African American youth (28%), and at least some representation of Latino youth (5%). However, because of the many studies not reporting ethnicity at all, it is difficult to know whether those figures are representative of most outcome research. A similar point can be made about socioeconomic factors. Some 72% of the studies provided no information of any kind about income or SES, thus limiting our ability to determine whether findings are relevant across the socioeconomic spectrum.

A closely related concern is the need to test moderators of treatment outcome. Many have stressed the need to identify moderators and to define the range within which beneficial treatments work, and have noted the failure of most studies to take this step (e.g., Durlak et al. 1995, Kazdin 2000, Kazdin et al. 1990, Weisz 2004). Our review suggests that the critical first step toward moderator assessment—i.e., collecting information on participant characteristics that might moderate effects—has not been taken in a remarkably large percentage of studies for some of the potential moderators that might be considered prime suspects and in which there is considerable societal interest (see previous paragraph). A related problem noted in our review is the high percentage of studies in which no data were reported for important variables such as who carried out the treatment and in what setting the treatment took place. Given the potential value of all these types of information to the field, and the lost opportunities for moderator assessment once data sets are no longer available, it seems desirable to encourage greater consistency in the kinds of information required by journals prior to acceptance of manuscripts for publication. If journal editors could agree on the contents, a simple checklist of information that must be included in the manuscript seems a relatively simple step to add to the review process (e.g., the Consolidation of the Standards of Reporting Trials statement for medication trials, Moher et al. 2001).

IDENTIFICATION OF PROBLEMS AND DISORDERS TO BE TARGETED IN TREATMENT PARTICIPANTS Turning to the procedures investigators used in their studies, we begin with one of the earliest and most critical steps in a clinical trial: determining whether potential participants do in fact have the problem or diagnosis that is to be targeted in treatment. Our figures (see Table 2) indicate that 75% of all the studies did not confirm their participant selection by obtaining a formal diagnosis for any individual in their sample, and that even among those that did take this step, about half did not use a reliable, standardized diagnostic assessment procedure. Instead, many used unstructured, nonstandardized approaches, including diagnoses based on clinician judgment, an approach that has been found to show little agreement with standardized diagnostic assessment (see, e.g., Jensen & Weisz 2002). Of course, a very reasonable alternative to diagnosis—a preferred alternative, in some respects—is the use of a cutoff on some standardized continuous symptom or problem measure, but our data show that more than 80% of studies did not take that step either. Given that more than 60% of studies did not report either a diagnostic procedure or a cutoff procedure for sample identification, it appears that we lack a precise picture of the target disorders or problems, or their severity, in most of the outcome evidence base. Indeed, it seems possible that a significant number of the youths treated did not actually meet criteria for the conditions targeted in the treatment.

The significance of this procedural gap in so many youth treatment outcome studies is highlighted by Chambless & Hollon's (1998) comment that ". . . if psychological treatment outcome research is to be informative, researchers must have clearly defined the population for which the treatment was designed and tested.

Thus, we do not ask whether a treatment is efficacious; rather, we ask whether it is efficacious for a specific problem or population" (p. 9). As Chambless & Hollon's comment implies, showing that a particular treatment had beneficial effects in a treatment trial may be of limited value if individuals in the trial were selected (*a*) via procedures of unknown validity, and (*b*) in ways that cannot be replicated in future applications. In the search for evidence-based treatments, a critical task is designing research in ways that permit matching of treatments to the groups for whom they have been shown to work. This matching process does not appear to be well supported in the evidence base to date.

RESEARCH DESIGN: SAMPLE SIZE Closely linked to the process of identifying youths who fit study criteria is the process of building a sample with adequate power to detect treatment effects. Chambless & Hollon (1998), citing Cohen (1988), have noted that for investigators to achieve 80% power to detect a medium difference between two groups in a treatment trial, sample size needs to be about 50 participants per condition. By these standards, the average youth psychotherapy trial has been substantially underpowered. As Table 6 shows, mean sample size across all trials in our review was 22 for treatment groups and 21 for control groups.

RESEARCH DESIGN: NATURE OF TREATMENT-CONTROL COMPARISONS Beyond statistical power, another factor that is closely linked to prospects for finding treatment effects is the type of treatment-control comparison investigators create in their studies. Our data (Table 6) show that the most common form of treatment-control group comparison employed in youth trials is the weakest experimentally: active treatment versus inactive waitlist or no treatment. Such passive control groups control only for the passage of time. Active control groups that control for time, attention, and nonspecific therapy/relationship factors were evident in fewer than 40% of the studies we reviewed, and standard case management as a control condition appeared in fewer than 8%. The one exception to this rather poor report card on treatment-control comparisons was the research base on ADHD treatment. For ADHD studies alone, active control groups involving attention and/or placebo interventions were more common than no treatment or waitlist control groups (70% versus 43%). It is possible, though speculative, that this reflects the close attention paid by ADHD treatment researchers to psychopharmacology trials (e.g., of stimulant medication), which must meet Food and Drug Administration standards; such trials often involve active treatment versus medication placebo conditions with participants blind to their condition.

RESEARCH DESIGN: NATURE OF THE TREATMENTS TESTED AND SUPPORT FOR TREATMENT INTEGRITY Our findings on the nature of the treatments that are compared to control conditions do nothing to mitigate the concerns of previous reviewers (e.g., Durlak et al. 1995, Kazdin et al. 1990) that behavioral and cognitive-behavioral treatments are disproportionately represented in the youth clinical trials literature. In studies featuring individual youth-focused treatments, 71% involved

learning-based interventions whereas only 8% involved insight-based approaches, broadly construed, and 9% involved eclectic approaches. A similar disproportion was evident in studies featuring parent-focused treatments and in studies employing family-focused treatments. Clearly, much ground remains to be covered in the future if the evidence base is to shed light on the impact of nonbehavioral interventions, which are in fact much more widely practiced outside research settings than the behavioral approaches that are so extensively tested.

Our review also characterizes the tested treatments in ways other than their theoretical basis. For example, we established that mean treatment duration was about 11 sessions, spanning about 9 weeks, and entailing about 13 hours of contact. These numbers varied considerably across the different problem clusters, with ADHD treatment averaging about 9 hours and conduct problem treatment about 18. Overall, use of group sessions (employed in 45% of studies) now appears to be as common as use of individual sessions (42%), and youths are the most common participants (involved in 83% of studies), with parents a distant second (25%). With so much attention given in current literature to the use of structured treatment procedures that involve participant homework (see, e.g., Chambless et al. 1996, Weisz 2004), it was interesting to note that only 40% of studies included any mention of homework assignments.

Finally, because the validity of any test of a treatment depends significantly on the faithfulness with which the treatment is delivered, we looked for evidence of fidelity support in study procedures. The results were not encouraging. Only 52% of studies reported using any form of a treatment manual, only 32% reported any pretraining of therapists, and only 32% reported either therapist supervision or adherence checks. Of course, this pattern could reflect incomplete reporting in some studies. However, it may also suggest that some of the research base entails tests of treatment models that were only weakly or erratically implemented by study therapists.

MEASUREMENT AND OUTCOME ASSESSMENT: INFORMANTS AND TECHNOLOGY EMPLOYED A somewhat more encouraging picture emerged from our focus on measurement. Although the youths participating in studies are popular as informants (used in 60% of all studies), significant percentages of the studies derived information from parents (33%), teachers (29%), and others (78%), including clinician raters and trained observers. The frequency with which self-report measures were used (66%) was almost matched by the frequencies for other-report measures (63%), objective behavior counts (63%), and behavioral observations by raters (58%, although the raters were blind to treatment condition of the youth in only 44% of the studies). In general, measurement strategies appear to draw from a healthy diversity of informants and to entail some use of respectable strategies for minimizing subjectivity and bias.

RANGE OF OUTCOMES ASSESSED FOLLOWING TREATMENT A significant challenge for treatment research is identification of the range and limits of treatment impact.

To meet this challenge may require outcome assessment that goes beyond the specific problems and disorders targeted by treatments. Indeed, in a conceptual model proposed by Hoagwood et al. (1996), a case is made for outcome assessment that includes not only target problems/symptoms and disorders, but also measures of such dimensions as real-world functioning, consumer perspectives, and impact on individuals and systems that relate to the treated youth. Given our criteria for the review, 100% of the studies included at least one measure of the target problem, symptoms, or disorder (we excluded many studies that did not measure what they set out to treat). In addition, 78% of all studies included at least one measure of nontarget problems or symptoms (e.g., a measure of depressed mood or self-esteem obtained in a study of ADHD treatment), 28% of all studies included at least one functioning measure (e.g., school grades or disciplinary incidents), 8% included at least one consumer satisfaction measure, and 5% included at least one measure of the environmental impact of treatment, broadly construed (e.g., parenting stress, assessed in a study testing behavioral parent training for child conduct problems). These numbers suggest that outcome assessment that transcends problems, symptoms, and diagnoses is not unheard of in the literature, but that such assessment is also not common practice.

CLINICAL REPRESENTATIVENESS OF STUDIES A final issue, about which our group has written frequently, is the clinical representativeness of study procedures. Our concern has been that a capacity to develop treatments that work well in actual clinical practice, and to determine how well they work in practice, may depend in part on the extent to which the characteristics of clinical trials resemble the characteristics of practice. Of course, clinical representativeness can involve multiple dimensions. Our analysis in this review has focused on three dimensions that seem particularly important—i.e., clinical representativeness of the youths sampled, the therapists who provide treatment, and the settings in which treatment is provided. Our findings suggest that the research base is not very clinically representative with respect to any of these three dimensions considered separately, and only 1% of the studies showed representativeness across all three dimensions—i.e., including at least some clinically referred children, some practicing clinicians, and some treatment in a clinical service setting. These limits on external validity of the evidence base illustrate the concern that has led to a proposed new model of intervention development and testing. This "deployment-focused model" (Weisz 2004, Weisz et al. 2004) acknowledges the need for initial efficacy evidence to establish the potential for treatment benefit. However, the model calls for such initial efficacy evidence to be followed in short order by research directed increasingly toward the kinds of individuals, interveners, and settings for which the intervention program is ultimately intended. The current state of the evidence base suggests that research conducted without reference to this model has used participants, therapists, and settings that depart markedly from those most germane to clinical practice.

CONCLUDING COMMENT

In conclusion, we stress that the evidence base shows some notable strengths but also some significant limitations. Our hope is that through periodic efforts to take stock of that base, investigators may identify ways in which refinement can contribute to the complementary challenges of improved intervention research and improved treatments for youths and families.

ACKNOWLEDGMENTS

We are grateful to Amie Bettencourt, Vickie Chang, Brian Chu, Jen Durham, Samantha Fordwood, Dan Fulford, Eunie Jung, and Robin Weersing for the important roles they played in this project, and to the John D. and Catherine T. MacArthur Foundation for its support of the project. The authors were also supported by the National Institutes of Mental Health (R01 MH 547347 and R01 MH 068806 to JRW, NRSA F31 MH65811 to AJD, and NRSA F31 MH12853 to KMH), by San Diego State University's Oscar Kaplan Postdoctoral Fellowship (KMH), and by the Department of Health and Human Services, Health Resources and Services Administration (Training Grant 1 D40 HP00017-01 to AJD), for which we express our thanks.

The *Annual Review of Psychology* is online at http://psych.annualreviews.org

LITERATURE CITED

Cantwell DP, Carlson GA, eds. 1983. *Affective Disorders in Childhood and Adolescence* New York: Spectrum

Casey RJ, Berman JS. 1985. The outcome of psychotherapy with children. *Psychol. Bull.* 98:388–400

Chambless DL, Hollon SD. 1998. Defining empirically supported therapies. *J. Consult. Clin. Psychol.* 66(1):7–18

Chambless DL, Sanderson WC, Shohan V, Bennett Johnson S, Pope KS, et al. 1996. An update on empirically validated therapies. *Clin. Psychol.* 49:5–18

Compton SN, Burns BJ, Egger HL, Robertson E. 2002. Review of the evidence base for treatment of childhood psychopathology: internalizing disorders. *J. Consult. Clin. Psychol.* 70(6):1240–66

Durlak JA, Fuhrman T, Lampman C. 1991. Effectiveness of cognitive-behavior therapy for maladapting children: a meta-analysis. *Psychol. Bull.* 110(2):204–14

Durlak JA, Wells AM, Cotten JK, Johnson S. 1995. Analysis of selected methodological issues in child psychotherapy research. *J. Clin. Child Psychol.* 24(2):141–48

Dush DM, Hirt ML, Schroeder HE. 1989. Self-statement modification in the treatment of child behavior disorders: a meta-analysis. *Psychol. Bull.* 106(1):97–106

Eysenck HJ. 1952. The effects of psychotherapy: an evaluation. *J. Consult. Psychol.* 16:319–24

Farmer EMZ, Compton SN, Burns JB, Robertson E. 2002. Review of the evidence base for treatment of childhood psychopathology: externalizing disorders. *J. Consult. Clin. Psychol.* 70(6):1267–302

Freud S. 1955. Analysis of phobia in a five-year-old boy. In *Standard Editions of the Complete*

Psychological Works of Sigmund Freud, Vol. 10, pp. 3–149. London: Hogarth

Hammen C, Rudolph KD. 1996. Childhood depression. In *Child Psychopathology*, ed. EJ Mash, RA Barkley, pp. 153–95. New York: Guilford

Hoagwood K, Jensen PS, Petti T, Burns BJ. 1996. Outcomes of mental health care for children and adolescents: I. A comprehensive conceptual model. *J. Am. Acad. Child Adolesc. Psychiatry* 35(8):1055–63

Jensen AL, Weisz JR. 2002. Assessing match and mismatch between practitioner-generated and standardized interview-generated diagnoses for clinic-referred children and adolescents. *J. Consult. Clin. Psychol.* 70(1): 158–68

Jones MC. 1924. A laboratory study of fear: the case of Peter. *Pedagog. Semin.* 31:308–15

Kazdin AE. 2000. *Psychotherapy for Children and Adolescents: Directions for Research and Practice*. Oxford, UK: Oxford Univ. Press

Kazdin AE, Bass D, Ayers WA, Rodgers A. 1990. Empirical and clinical focus of child and adolescent psychotherapy research. *J. Consult. Clin. Psychol.* 58(6):729–40

Lefkowitz M, Burton N. 1978. Childhood depression: a critique of the concept. *Psychol. Bull.* 85:716–26

Levitt EE. 1957. The results of psychotherapy with children: an evaluation. *J. Consult. Clin. Psychol.* 21:189–96

Levitt EE. 1963. Psychotherapy with children: a further evaluation. *Behav. Res. Ther.* 60:326–29

McLeod BD, Weisz JR. 2004. Using dissertations to examine potential bias in child and adolescent clinical trials. *J. Consult. Clin. Psychol.* 72:235–5l

Moher D, Schulz KF, Altmann DG. 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Med. Res. Method.* 1:2

Prout HT, DeMartino RA. 1986. A meta-analysis of school-based studies of psychotherapy. *J. Sch. Psychol.* 24(3):285–92

Rochlin G. 1959. The loss complex. *J. Am. Psychoanal. Assoc.* 7:299–316

Shirk SR, Russell RL. 1996. *Change Processes in Child Psychotherapy: Revitalizing Treatment and Research.* New York: Guilford

US Department of Health and Human Services. 2001. *Mental Health: Culture, Race, and Ethnicity—A Supplement to Mental Health: A Report of the Surgeon General*. Rockville, MD: US Dept. Health Hum. Serv., Subst. Abuse Mental Health Serv. Admin., Cent. Mental Health Serv.

Weisz JR. 2004. *Psychotherapy for Children and Adolescents: Evidence-Based Treatments and Case Examples.* Cambridge, UK: Cambridge Univ. Press

Weisz JR, Hawley KM. 1998. Finding, evaluating, refining, and applying empirically supported treatments for children and adolescents. *J. Clin. Child Psychol.* 27(2):206–16

Weisz JR, Jensen AL, McLeod BD. 2004. Milestones and methods in the development and dissemination of child and adolescent psychotherapies: review, commentary, and a new deployment-focused model. In *Psychosocial Treatments for Child and Adolescent Disorders: Empirically Based Strategies for Clinical Practice*, ed. ED Hibbs, PS Jensen. Washington, DC: Am. Psychol. Assoc. 2nd ed. In press

Weisz JR, Weiss B, Alicke MD, Klotz ML. 1987. Effectiveness of psychotherapy with children and adolescents: a meta-analysis for clinicians. *J. Consult. Clin. Psychol.* 55(4): 542–49

Weisz JR, Weiss B, Han SS, Granger DA, Morton T. 1995. Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. *Psychol. Bull.* 117(3):450–68