Journal of Clinical Child Psychology
1998, Vol. 27, No. 2, 206–216

Finding, Evaluating, Refining, and Applying Empirically Supported Treatments for Children and Adolescents

John R. Weisz and Kristin M. Hawley
Department of Psychology, University of California, Los Angeles

Structured child and adolescent treatments, tested through controlled clinical trials, have produced beneficial effects in hundreds of studies. By contrast, the limited pool or research on traditional clinical treatments raises doubts about their effectiveness. Thus, identification of empirically supported treatments may contribute something of real value to clinical practice and training. The Child Task Force report represents an important initial step in this direction. Here we offer both praise and critique, suggesting a number of ways the task force process and product may be improved. In addition, we suggest several ways to strengthen and enrich the clinical trials research available to the Task Force, emphasizing the need to test empirically supported treatments with referred youth in practice settings.

After many years of planting, nurturing, cross-pollinating, and cultivating, clinical researchers have brought forth a promising crop of treatments for children and adolescents (herein referred to collectively as *children*). The American Psychological Association Task Force (Task Force on Promotion and Dissemination of Psychological Procedures, 1995) charged with harvesting that crop has been busy, and to good effect. Those of us in the field who have worked to strengthen the connection between clinical research and clinical practice will find much to appreciate in the Task Force report and in the process this report represents. It seems clear that if we are to strengthen ties between research and practice, one essential step must be identification of those products of sound research that have potential for clinical use. In this regard, the Task Force report on child treatments can contribute importantly. On the other hand, this report is but one early step in what is likely to be an ongoing and often complex process. The report does not, and of course could not, answer all the questions that will ultimately need to be addressed. This premise of necessary incompleteness underlies much of the present article, particularly those parts in which we look to the future.

We begin this article by viewing the Task Force report from a meta-analytic perspective, commenting on relevant findings from quantitative reviews of child psychotherapy outcome research. Then we address the interface between meta-analytic findings and the Task Force work. Next we note both strengths and limitations of the Task Force work, and we offer recommendations for next steps in the process. Finally, we offer a critique of child psychotherapy outcome research in general, noting several ways that we think the field might be improved and addressing issues related to the exportability of research-derived treatments to clinical practice settings.

Preparation of this article was facilitated by a National Institute of Mental Health Research Scientist Award K05 MH01161 and Research Grants R01 MH49522 and R01 MH57347, which we gratefully acknowledge.

Requests for reprints should be sent to John R. Weisz, Department of Psychology, Franz Hall, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095–1563. E-mail: weisz@psych.ucla.edu

Meta-Analyses: Caveats, Findings, and Relevance to the Task Force Work

There are a number of ways to survey and summarize what we know about treatments for children and how well they work. Meta-analyses (see Mann, 1990) are a useful complement to the Task Force approach. Here we describe and critique meta-analyses in the child area, and we note some particularly relevant findings.

Overview of Meta-Analysis

In psychotherapy meta-analyses, a common effect size (ES) metric is applied to a collection of treatment outcome studies to permit pooling of findings across the studies. In most meta-analyses of child treatment research, the ES is the difference between posttreatment (or follow-up) means for treated and untreated youth on an outcome measure of interest, divided by the standard deviation of the measure. Computing the ES mean for any treatment group versus control group comparison typically involves averaging ES values across the multiple outcome measures used in a study.

By pooling these average ES values across studies, meta-analyses can generate estimates of overall treatment impact; compare outcomes among theoretically meaningful subsets of studies; and test child, therapist, and therapy characteristics that may moderate treatment outcome.

Meta-Analysis: Caveat Emptor

Although these estimates, comparisons, and tests can be useful heuristically and as summaries of the state of the evidence, their proper interpretation requires attention to some of the limitations of meta-analysis. First, as with any summarizing technique, the output of meta-analysis inevitably reflects the limitations of input. Two examples illustrate the point. First, more than 75% of the studies in most child meta-analyses test behavioral (including cognitive-behavioral) interventions; thus, current meta-analyses simply cannot provide as representative a picture of psychodynamic and other nonbehavioral treatments as they can behavioral approaches. A second example is that meta-analyses have thus far been applied to studies involving between-group comparisons, with a large body of evidence from within-group designs (e.g., Pelham et al., 1993) and single-subject designs (e.g., Tarnowski, Rosen, McGrath, & Drabman, 1987) omitted.

A second limitation is the unavoidable confounding among factors that may relate to outcome. For example, particular kinds of treatment tend to be associated with particular child problems; thus, main effects of treatment type will reflect, in part, main effects of type of treated problem. This problem can be addressed in part through eliminating and interaction tests (see e.g., Weisz, Weiss, Alicke, & Klotz, 1987), but given the large number of factors that may relate to outcome, it would be unrealistic to think the problem could be fully solved. Finally, an array of potential limitations relates to the dozens of methodological decisions that must be made in any meta-analysis. Examples include such decisions as the following:

1. How rigorous must studies be to merit inclusion? For example, must children have been randomly assigned to treatment and control groups?
2. Should ES calculation be based on only "hard measures" (e.g., behavior counts, success in approaching a previously feared object), or should "softer measures" such as subjective ratings be included?
3. Should outcome measures only be accepted if they come from informants who are blind to participants' treatment condition? (This seems a good idea in principle, but it would rule out reports by both child participants and their parents, who are arguably the two most knowledgeable informants.)
4. Should ES values be averaged in raw form or weighted according to sample size?

Our research suggests that these and numerous other methodological decisions may relate to the magnitude of ES values generated in a meta-analysis[1] (see Weiss & Weisz, 1990; Weisz, Weiss, Han, Granger, & Morton, 1995). Of course no two meta-analytic teams will make all methodological decisions in the same way. An important consequence is that differences between the findings of any two meta-analyses will reflect method variation, in addition to truly substantive differences in the magnitude of treatment effects in the studies sampled. This state of affairs makes convergent findings across different meta-analyses particularly noteworthy, and findings of the major meta-analyses of child treatment research have been quite convergent (see Weisz & Weiss, 1993).

Meta-Analytic Findings

We know of four broad-based child psychotherapy meta-analyses—that is, meta-analyses involving diverse collections of studies, with few limits imposed on the kinds of treated problems or types of intervention that are included. Together, these four meta-analyses encompass more than 300 separate treatment outcome studies. In the first of the four, Casey and Berman (1985) included outcome studies published between 1952 and 1983, focusing on treatment of children age 12 and younger. The mean ES was 0.71 for those studies that included treatment–control comparisons; as an aid to interpretation, ES values of .20 have been regarded as "small," .50 as "medium," and .80 as "large" (guidelines derived from Cohen, 1988). Viewing the Casey–Berman findings in percentile terms, the average treated child in the studies they surveyed scored better after treatment than 76% of control group children, averaging across outcome measures. In a second meta-analysis, Weisz et al. (1987) reviewed outcome studies published between 1952 and 1983, involving children ages 4 to 18. The mean ES was 0.79, indicating that, after treatment, the average treated child was at the 79th percentile of control group peers across outcome measures. In another meta-analysis, Kazdin, Bass, Ayers, and Rodgers (1990) included studies published between 1970 and 1988 with children ages 4 to 18. For the subset of studies that compared treatment and no-treatment control groups, the mean ES was 0.88; the average treated child scored higher, after treatment, than 81% of the no-treatment comparison group. For studies in the Kazdin et al. collection that involved comparison of treatment groups to active control groups, mean ES

[1] In general, we have found that methodological rigor in outcome studies is *positively* correlated with ES; this suggests that current meta-analyses, which include studies across a range of methodological sophistication, may actually underestimate the true magnitude of treatment effects (see Weiss & Weisz, 1990).

was 0.77; the average treated child functioned better, posttreatment, than 78% of the control group. The fourth broad-based meta-analysis, by Weisz, Weiss, et al. (1995), included studies published between 1967 and 1993, involving children ages 2 to 18. The mean ES of 0.71 meant that, after treatment, the average treated child was functioning better than 76% of comparison children in the control groups. (For more detailed descriptions of the procedures and findings of various meta-analyses, see Weisz & Weiss, 1993.) These four broad meta-analyses present a consistently positive picture, with a mean ES not far below the 0.80 level used to indicate a "large" effect.[2]

Complementing the four broad-based child meta-analyses are others that tackle rather specific questions by focusing on select subsets of child outcome studies. Meta-analyses confined to cognitive-behavioral therapy have found substantial positive effects across a range of target problems (Durlak, Fuhrman, & Lampman, 1991) and on impulsivity considered alone (Baer & Nietzel, 1991). Also, Dush, Hirt, and Schroeder (1989) found significant positive effects associated with the specific cognitive-behavioral technique of self-statement modification. Two meta-analytic teams (Hazelrigg, Cooper, & Borduin, 1987; Shadish et al., 1993) found beneficial effects of family therapy. Moderately positive effects have been found for interventions used to prepare children for medical and dental procedures (Saile, Burgmeier, & Schmidt, 1988) and for psychotherapy administered in school settings (Prout & DeMartino, 1986). Finally, the broad range of questions to which meta-analysis may be applied is illustrated by Russell, Greenwald, and Shirk's (1991) test of whether child language proficiency improved with psychotherapy across a sample of relevant studies; it did, particularly with therapies emphasizing spontaneous verbal interaction.

Meta-Analytic Findings, Clinic Therapy Findings, and the Task Force Mission

In several respects, the various meta-analyses offer encouragement that the field of child psychotherapy is sufficiently mature to warrant serious task force attention. First, the verdict of the four broad-based meta-analyses, summarizing more than 300 independent outcome studies, is uniformly positive. Findings show a substantial overall mean effect that lies well within the range of what has been found for adult psychotherapy (see, e.g., Shapiro & Shapiro, 1982; Smith & Glass, 1977). Second, recent evidence indicates that treatment effects have substantial specificity. That is, effects are markedly stronger for the precise problems the treatments are designed to address than for nontargeted problems (Weisz, Weiss, et al., 1995). Third, follow-up tests, averaging about 6 months after the treatments have ended, show ESs that are remarkably similar to immediate posttreatment effects (see Weisz et al., 1987; Weisz, Weiss, et al., 1995), suggesting that child treatments are producing durable effects, at least within the time frame of the typical assessment.[3]

One other body of evidence points to the potential value of the Task Force work: findings on the effectiveness of child treatment in everyday clinical practice. The clientele, therapists, settings, and procedures of most conventional clinical practice differ substantially from those of most clinical trials summarized in meta-analyses. Noting these differences, and prompted by our own findings on this issue, we searched (in Weisz, Donenberg, Han, & Kauneckis, 1995; Weisz, Donenberg, Han, & Weiss, 1995; Weisz, Weiss, & Donenberg, 1992) for acceptably designed studies that (a) involved treatment of clinic-referred children, (b) with treatment carried out in service-oriented clinics or agencies, (c) with therapy conducted by practicing clinicians, and (d) with treatment done as part of the usual service delivery function of the clinic (i.e., not primarily for a research project). Our most recent search (Weisz, Donenberg, Han, & Weiss, 1995) netted only nine such studies. ES estimates for these nine studies suggested generally negligible effects. Recent evidence has also cast doubt on the effectiveness of the "system of care" approach, in which multiple conventional child treatments and services are provided to children through the assistance of case managers (see, e.g., Bickman, 1996; Weisz, Han, & Valeri, 1997; Weisz, Walter, Weiss, Fernandez, & Milkow, 1990). If further studies should continue to indicate that conventional clinical treatments are not very effective—and further studies are certainly warranted—then the need to identify beneficial treatments would be particularly urgent. That is, to the extent that treatments used in current clinical practice fall short of the desired effects, the importance of the Task Force search for beneficial treatments is magnified.

In summary, meta-analyses have shown that empirically tested child treatments, on average, produce substantial positive effects that are (a) similar in magnitude to the effects of psychotherapy with adults, (b) relatively specific to the problems targeted in treatment, and (c) relatively lasting in their impact. And companion analyses of the limited pool of evidence on conventional clinical treatments have thus far generated a discouraging picture of their effectiveness, thus suggest-

[2]Note, however, that recent analyses (in Weisz, Weiss, et al., 1995) suggest that true population ES means, adjusting for heterogeneity of variance, may be closer to the "medium" level.

[3]This optimistic note is tempered somewhat by the fact that fewer than half the studies in our two meta-analyses reported any follow-up assessment.

ing that there may be a need to identify new, empirically supported treatments. Thus, taken together, the meta-analytic research and the companion findings on clinical treatment provide a supportive rationale for the Task Force mission.

Strengths and Limitations of the Task Force Work

A case can be made not only for the potential value of the Task Force enterprise but also for several aspects of the approach used by both the child and adult groups. The approach also has limitations, in our view; some of these warrant close attention because they suggest possible future directions for the Task Force.

Some Strengths of the Task Force Approach

The thoughtful approach taken by the Task Force has many strengths. We offer three illustrations. First, requiring manualization (or other means of insuring faithful replication of treatment procedures) serves the important objective of making the identified treatments available for clinical training and practice. Second, requiring clear specification of sample characteristics reflects increasing recognition that specific treatments may be efficacious only within a limited range of such person characteristics as age, problem severity, and ethnicity (see Weisz, Huey, & Weersing, 1998). And third, requiring that supportive findings for a treatment be replicated by an independent research team before that treatment can be designated as "well established" sets an admirable standard.

Limitations of the Task Force Work and Challenges for the Future

The Task Force reports are perhaps best viewed as works in progress, subject to ongoing improvements and refinements. From this perspective, it is worthwhile to note that potential limitations of the current approach may need attention in future work.

Multiple, subjective judgments that are not yet well-specified or standardized. One of the lessons learned from meta-analytic work is that qualitative judgments about treatment outcome studies can differ greatly from one reader to the next. This fact necessitates the development of uniform standards and guidelines, with coding manuals and reliability assessment. This problem has not yet been well-addressed by either the adult or child Task Force. Current procedures require an extensive list of qualitative, subjective judgments for which relatively few specific criteria have been provided, few coding rules developed, and no interjudge reliability assessment conducted. This is a risky state of affairs, in our view.

For example, to qualify as empirical support under Task Force Criteria 1 and 2a (see Lonigan, Elbert, & Johnson, this issue), studies are required to have either "well-conducted group design" or "good [single case] experimental designs." Whereas many in the field would no doubt favor good designs over bad, there might be considerable disagreement about how to define the terms. To qualify as "well conducted," must a between-group study have employed random assignment (and with or without group-matching procedures?), used objective (i.e., performance test) outcome assessment or blind informants, assessed treatment integrity, or achieved similar and low attrition rates in treatment and control groups? Failure to meet any one of these criteria (and many others, as well) could threaten validity, but which failure, or combination of them, if any, should mean that a study cannot be cited as support for a treatment program? We suspect that different Task Force members would answer this question in different ways. With members making their judgments about candidate studies independently of uniform standards and relatively independently of one another (i.e., with one member surveying depression treatment studies, another ADHD studies, etc.), the risk is that the Task Force judgments are relatively unreliable. To fill out the picture, a few additional examples may be useful. Criterion 1b (Lonigan et al., this issue) holds that treatments may be regarded as supported if they prove "equivalent to an established treatment," but it is not clear how raters should decide what treatments are "already established." Criterion 4 (Lonigan et al., this issue) specifies that "sample characteristics must be clearly specified;" as noted earlier, we agree with this general notion, but we must add that it is not clear *which* characteristics must be specified, nor which kinds of omission should mean that a study cannot be cited as support for a treatment. Criterion 3 (Lonigan et al., this issue) requires that studies employ treatment manuals; we agree with this general principle, as previously noted, but we have seen the term *manual* used in diverse ways in the child area and applied to documents ranging from a few pages in outline form to more than 300 pages of very detailed instructions. Must Task Force members see the manual before classifying the study, and if so, what criteria should be applied to judge whether Criterion 3 has been satisfied?

In a recent paper, Chambless and Hollon (1997) offered many valuable ideas on how various judgments should be made in the process of evaluating studies. These ideas, together with other statements on guidelines (e.g., Chambless et al., 1996; Lonigan et al., this

issue; Task Force on Promotion and Dissemination of Psychological Procedures, 1995), certainly provide a strong nucleus for what could become a uniform set of standards for Task Force judgments about studies and treatments.

Defining a *treatment*. Two closely related concerns require separate attention because they are so basic to the Task Force mission. The first is the matter of how a *treatment* should be defined. In our own reviews of empirically supported treatments, we have found that many investigators, in efforts to improve their treatments, revise their manuals from one study to the next, in some cases producing quite different versions of the treatment from one incarnation to the next. We find no guidelines in the Task Force documents for deciding when two different versions of a treatment program should and should not be considered the same treatment. In an additional complication, investigators frequently describe their tested treatment as based "partly" or "largely" on a treatment developed by another investigator. Here too, the Task Force appears to have no specific guidelines for deciding whether two separate treatments are the same. Chambless et al. (1996) did note that in defining interventions, "brand names are not the critical identifiers. The manuals are" (p. 6). Some variability in manuals across studies seems inevitable. What remains unclear is just how similar manuals must be to qualify as representing the same treatment.

To illustrate why this issue is important, we note two examples in the child area. First, Barrett, Dadds, and Rapee (1996) found significant effects of their 12 session, Australian "Coping Koala" adaptation of Kendall's 16- to 20-session "Coping Cat" Child Behavior Therapy (CBT) program for child anxiety (see Kendall, 1994), relative to a wait-list control condition. The Barrett et al. program was clearly designed to be similar to Kendall's program, but there are content differences, and the Australian version is 4 to 8 sessions shorter. Whether the two are judged "the same" has immediate implications for the Task Force process. If the Barrett et al. intervention and the Kendall intervention are classified as the same treatment, then we have a single-treatment program meeting criteria for the "well-established" Task Force category, because it has been supported by two different investigative teams in two very different locations. If the two interventions are classified as different treatments, then neither qualifies as well established.

A similar but more complex situation arises with regard to treatment of depression in children and adolescents. Recent reviews by Kaslow and Thompson (this issue) and Weisz, Valeri, McCarty, and Moore (in press) indicate that there are at least six studies showing significant effects of multifaceted CBT packages. The overlap among the packages varies, with most including such venerable components as pleasant activity selection, cognitive retraining, and social skills training, but each treatment has distinguishing features, and foci, different from the others; moreover, the age groups treated differ across studies from middle childhood to late adolescence, with corresponding structural and content differences in the respective manuals. Following a liberal interpretation, it might be argued that we have one treatment, "CBT for child and adolescent depression," which is well established because it has been supported in six studies by different teams. Following a conservative interpretation, we may have six different CBT treatments, each supported in only one study. The risk, of course, is that different Task Force members will follow different conventions if no uniform guidelines are provided, and thus whether a particular treatment is or is not classified as "empirically supported" will depend on which Task Force member makes the judgment.[4]

Defining *empirical support*. Another basic concern applies to the seemingly straightforward Task Force task of identifying studies "demonstrating efficacy" (Chambless et al., 1996). Chambless et al. did note that the Task Force has focused on "tests of change in the defining problem or symptoms" (p. 6; i.e., not on problems that were not the primary target of treatment). But what judgment should be made about a study that (a) demonstrates efficacy on two out of five of these target outcome measures, (b) shows significant effects only on participant self-report measures but not on more objective measures, (c) shows a significant treatment versus wait-list difference at posttreatment but flunks all tests of clinical significance, or (d) shows effects at immediate posttreatment but not at 2-month follow-up? Outcome studies rarely show uniform results across all measures and all assessment points; the Task Force currently lacks uniform standards for judging which combinations of outcomes should be viewed as demonstrating efficacy and which should not.

Lack of a common outcome metric across studies—uneven playing field. This last point is related to another general concern about the Task Force process for classifying studies and treatments: No common outcome metric is applied to the studies. Instead, treatments are judged "well established," "probably ef-

[4]One other point bears attention here: If the Task Force were to adopt conservative standards, defining "same treatment" as "exactly the same manual," then our field would need much more precise replication research than is currently being conducted. Indeed, under this strict interpretation, there may be virtually no true replication in the field, at present.

ficacious," or neither on the basis of traditional statistical tests of group differences. Results of such tests are influenced by sample size and, even under the best of circumstances, provide only a gross index of treatment efficacy. For any treated condition, the pool of studies showing "empirical support" in the form of significant *p* values may encompass a very broad range of effect magnitude; under Task Force procedures, significant effects are classified as equally supportive regardless of magnitude. Consequently, two studies showing significant *p* values and ESs of, say, .40 and 1.60, respectively, can contribute equally to the standing of their treatment program under current Task Force procedures despite a fourfold difference in magnitude of benefit. In fact, current procedures make it quite possible for studies with trivial effects to be classified as supportive if they employ large samples and generate statistically significant *p* values, even while studies with *larger* ESs but smaller samples, and thus nonsigificant effects, are classified as unsupportive. Given such concerns, the Task Force may need to consider supplementing significance test findings with ES information, thus providing more comparable standards for evaluation across studies.

How to use or weight negative and null findings. Even if a common metric were employed, the Task Force would continue to confront another question: What can be done with studies that report no effect of a particular treatment program and with those few that show negative effects indicating worse outcomes for treated participants than for control groups? As best we can tell, the Task Force work thus far has focused on tallies of supporting studies, with no established procedure for taking account of contrary evidence. It was good to see this issue addressed in the recent statement by Chambless and Hollon (1998). They argued that conflicting findings should lead to a focus on the quality of the conflicting studies and that a treatment for which there is truly mixed evidence should not be in the Task Force list until the factors leading to different outcomes across studies are understood. These general principles could provide a reasonable starting point for development of consistent procedures and guidelines for Task Force reviewers.

Enriching the Task Force product. Implicit within much of this section is the possibility that the Task Force reports might provide more information of value to clinicians, training programs, and others than is currently provided by indicating only whether treatments are well established or probably efficacious. For example, users of the report might find useful some indication of the *magnitude* of effects generated in the relevant studies, the *methodological quality* of those, the *reference groups* (e.g., age range, ethnicity, problem severity) with which the treatment has been tested, and the degree to which the treatments have been tested in *clinically representative conditions* (see next). These and many other kinds of information could be valuable to training programs and clinicians, as well as researchers. How far beyond the current two-category system the resources of the all-volunteer Task Force will permit the group to move remains to be seen.

What Our Discipline Can Do to Support the Task Force Mission

For the Task Force to do its best work and maximize its value to others, our discipline may need to make certain adjustments. Three examples are noted here.

More Consistent Reporting Across Studies and Across Journals

If the Task Force were to move in the directions that we previously suggested, much more uniform reporting would be needed in clinical trials articles than is currently the case. For example, if the Task Force review were to take account of study methodological quality in some systematic way, information might be needed for all studies on such questions as (a) whether the groups were randomly assigned, and with or without subsequent matching on critical dependent variable; (b) the source reliability and validity of the outcome measures employed, and which measures, if any, came from informants blind to participants' treatment condition; (c) results of integrity tests gauging the degree to which therapy sessions adhered to the manual; and (d) attrition rates in all groups, at all assessments after treatment.

If the Task Force were to make classification of studies into its categories contingent on the kinds of information now listed as important in its guidelines, then study authors might need to include in their reports (a) specific details on the length and content of any manuals involved, not simply statements that a "manual" was used; (b) exact figures on sample characteristics (e.g., age, sex, ethnicity) by group; and (c) any of numerous kinds of descriptive information noted in reports by the Task Force on Promotion and Dissemination of Psychological Procedures (1995), Chambless et al. (1996), and Chambless and Hollon (1998). Finally, if the Task Force were to move in the direction of a common metric for effect magnitude, as we previously suggested, then members would need more than mere tests of statistical significance; instead, critical information would include means and standard deviations for all study groups at all points of assessment and perhaps ESs for each post-treatment and follow-up group comparison. Move-

ment toward more explicit criteria for Task Force member judgments could be nicely supported by journal editorial policies that create consistent reporting for clinical trials.

Increased Support for Replication

The premium that the Task Force has placed on replication of findings conflicts to some degree with the emphasis in our discipline on originality of contributions. A broad range of decisions in academia, ranging from hiring to tenure to grant funding, are tilted to favor original contributions over repeat performances and retests of other investigators' findings. The Task Force has rightly stressed our need to know whether outcome findings are robust or mere flashes in the pan, but to satisfy this need, our field may need to find ways to encourage, finance, and reward carefully done replications. Otherwise, the incentive system in the discipline may continue to work against an important form of research that remains essential to full evaluation of treatments.

Testing Empirically Supported Treatments Under Real-World Conditions

The Task Force was created, in large part, to identify a list of treatments so well supported that they might warrant use in clinical practice. However, most of the child treatments identified thus far (and possibly most of the adult treatments, as well) have been tested neither in conventional clinics nor under conditions that very much resemble clinical practice. As a consequence, we actually know little about whether these treatments will be effective in clinical use, despite the fact that they have empirical support. Numerous differences between the characteristics and conditions of therapy in clinical settings and most therapy in treatment outcome research have been detailed elsewhere (e.g., Weisz et al., 1992; Weisz, Donenberg, et al., 1995). A few of the differences, and their implications, need special emphasis here.

Client severity and motivation. A majority of the clinical trials for children have tended to focus treatment on recruited youths, most of whom have not been shown to qualify for a diagnosis and most of whom were not seriously disturbed enough to have been referred for treatment had the study not been conducted. In addition, because clinical trials must meet human participants research requirements, all the participants are study volunteers who are uncoerced and motivated to receive the treatment. By contrast, our research in community clinics suggests that child motivation for treatment is low, relative to the parents who typically initiate treatment, and that even a significant percentage of the parents initiated treatment only when coerced by police, court, or child protective services. It is not clear how treatments tested with motivated study volunteers who are not seriously disturbed will fare in clinics where the clients are seriously disturbed and not nearly so motivated. Hence, we need more outcome research evaluating empirically supported treatments with the kinds of children and families clinicians in the real world are most likely to encounter. Such research can help us learn whether the manualized treatments are effective without change or require adjustments (e.g., to address higher levels of severity and/or lower levels of motivation).

Exclusionary criteria. Clinical trials researchers frequently seek optimal candidates for a particular treatment. So, they set exclusionary criteria, ruling out, for example, single-parent or foster care children, children with substance use problems, and so forth. Children with such complicating characteristics are more likely to end up in real clinics, not clinical trials. Are the researchers correct in believing that these complicating factors will undermine treatment success? If so, then the treatments they have developed in the absence of such factors may need retooling if they are to be effective in real-world clinical practice, where exclusions are often disallowed.

Comorbidity. In research we are currently conducting in child outpatient community clinics (Weisz et al., 1998), standardized diagnostic interviews point to a mean of more than 3.5 *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., rev.; American Psychological Association, 1987) diagnoses per child. The level of comorbidity in such clinical settings is almost certainly higher than in the typical child clinical trial, where many of the children have no diagnosis and where significant comorbidities are sometimes ruled out of the sample. Taking lab-tested treatments into clinical settings is apt to mean taking on more substantial comorbidity, on average, than the treatments have yet confronted, and some of the comorbid conditions certainly may complicate treatment; the difficulty of treating depression, for example, may increase markedly when the condition is combined with conduct disorder. Of course, we may find that existing treatments work well with many forms of comorbidity, perhaps even producing benevolent effects on the comorbid conditions. Alternatively, we may find that many current manuals need modification, perhaps with branching optional modules designed to address particular comorbid problems sufficiently to prevent them from undermining treatment. Or, conceivably, we will learn

that some of our treatments that work well with pristine and uncomplicated cases are simply not very effective with certain combinations of child dysfunction, even with additional modules added. Such findings might prompt a new generation of treatment manuals designed to address specific combinations of disorders that frequently occur in concert—for example, a manual for co-occurring depression and anxiety, another for co-occurring depression and conduct disorder. Of course, none of these steps will occur until we extend our treatment research into the murky terrain of comorbidity.

Lockstep manuals versus the unpredictability of real life in clinics. Treatment manuals we have seen in the child area rarely involve fewer than 8 sessions, and most are considerably longer, with some manuals calling for more than 24 sessions; this applies to treatment of both internalizing (e.g., Stark & Kendall, 1995) and externalizing conditions (e.g., Lochman, Coie, Underwood, & Terry, 1993). It is also in the nature of most manuals to build knowledge and skills in a logical, cumulative sequence, such that skill acquisition in any one session may well depend on the base built in previous sessions. Most children in the treatment groups of most published clinical trials complete most of the manualized sessions; those who complete too few to cross the completion threshold are generally excluded from analyses. Thus, the positive outcomes reported in most clinical trials often reflect only the experience of successful treatment completers.

This seems reasonable, from a research perspective, but it is a cause for some concern when weighed against treatment and termination patterns in many conventional clinics. Our research in community clinics (e.g., Weisz & Weiss, 1989) suggests two worrisome conclusions: (a) In many outpatient clinics, the mean number of treatment sessions completed before termination is fewer than eight and thus less than the shortest of empirically supported treatment manuals, and (b) treatment termination is very frequently initiated by the child and family and very often without prior announcement (i.e., they simply stop coming to the clinic). In other words, children often stop coming after relatively few sessions and in ways that rule out an opportunity for therapists to summarize, highlight, or achieve real closure. What is not clear from empirical research to date is how our lengthy, lockstep, cumulative manuals will fare in the face of these hard clinical realities. Is completion of half a manual, or less, without ever reaching final closure going to improve outcomes for children who drop out? It is possible that manuals designed for use in real clinical practice will need to differ structurally or substantively from their lab-tested counterparts (e.g., with "bunching" of key treatment lessons very early in treatment) to ensure that even early dropouts will have been exposed to the most basic ideas before they stop attending. Alternatively (or in addition), we may find that the organization and structure of the manualized approach motivates children and parents to attend clinic sessions more faithfully than is the case with less structured, traditional approaches. Lessons of either type would be valuable in guiding the exportation of manualized treatments from research settings to the practice settings where most of the real treatment actually takes place.

Developing effectiveness tests for use by the Task Force. To address issues like these, we need an array of strategies for effectiveness research, including (a) tests of manualized treatments with clinically referred youths, in clinic settings, with clinic-employed therapists doing the treatment; (b) tests of structured treatments based on models widely used in clinical practice but largely ignored in most current outcome research; and (c) comparison of outcomes of usual care in clinics with outcomes for comparable clients treated with empirically supported treatments. These are but a few of dozens of ways that we researchers could make our work much more directly relevant to the lives of clinical practitioners and trainers than is currently the case. As research on this theme accumulates, it may be worthwhile for Task Force members to consider an additional dimension along which treatments may be rated: readiness for clinical use. Such a rating would reflect the degree to which a treatment has been found effective under circumstances (e.g., clients, therapists, settings) resembling actual clinical practice.

Strengthening Research on Child and Adolescent Treatment

There are several other ways treatment researchers can help to advance our field, both broadening and deepening the array of evidence available to the Task Force. We conclude by noting some particularly relevant examples.

Enriching Outcome Assessment

Most child outcome studies to date have emphasized measures of symptoms and diagnosis, often with relatively little emphasis on whether treated youth improve on functional dimensions that matter most to them. We need to know much more about whether our treatments improve daily life functioning in such critical settings as home, school, and peer group. Indeed, if our treatments are to move from research laboratories to clinics, we will need attention to not only symptoms and func-

tioning, but to other dimensions noted in Hoagwood, Jensen, Petti, and Burns's (1996) recent model of child mental health outcomes: consumer perspectives (e.g., satisfaction with treatment), environments (e.g., changes in family relationships), and systems (e.g., subsequent use of mental health services).

Lengthening Outcome Assessment

A majority of child treatment outcome studies, to date, have not provided any treatment-control comparison on outcome measures beyond the immediate posttreatment assessment. For the minority of studies that have reported follow-up tests, the mean lag from termination is about 6 months (see Weisz & Weiss, 1993). Clearly, we need to know much more about the time course of treatment effects than this limited base of information affords. Some treatments that now seem impressive may not be found to produce lasting effects; learning the limits might prompt research on booster sessions and periodic checkups to assess slippage. Other treatments may take time to "kick in," especially for the kinds of real-life, outside-therapy outcomes discussed in the previous paragraph (e.g., grades, family relationships), some of which may show more gradual treatment effects. Moreover, because child treatment occurs during a period of rapid cognitive change, cognitive development may interact with treatment inputs, but the results of this interplay might not be evident within the 6-month mean lag of most follow-up assessments. We would not want to unfairly conclude that a treatment program is ineffective simply because we terminated outcome assessment before the full impact of the treatment had taken shape. So, for several reasons, we need to begin building a base of information on longer term outcomes of our interventions.[5]

Assessing Moderators of Treatment Outcome

We also need to construct a picture of the range of child and family characteristics within which particular treatments are beneficial. Full descriptions of sample characteristics will certainly help, but we also need much more attention by researchers to variables *within* their samples that moderate outcome. To focus on just a few examples, we note that nearly all child researchers have ready access to such basic information as child age, sex, and ethnicity, but direct tests of whether any of these factors relates to outcome are exceedingly rare in the outcome literature. Meta-analyses (see e.g., Durlak et al., 1991; Weisz, Weiss, et al., 1995) have shown both age and sex effects, plus an Age × Sex interaction, averaging across studies, but parallel within-study analyses are rare. As for ethnicity, about 80% of child outcome studies have failed to even report the composition of their samples (see Kazdin et al., 1990), and when the statistics are reported, little seems to be done to evaluate their significance for outcome. In a recent search, Weisz et al. (in press) found 19 child and family outcome studies in which authors reported ethnicity and in which the majority of the sample were ethnic minorities; only 1 of the 19 reported any direct test of whether outcomes were different for different ethnic groups. The limited attention given to even such obvious candidate moderators as age, sex, and ethnicity suggests that our picture of child and family factors that may influence outcome has barely begun to be drawn.

Assessing Mediators of Treatment Outcome: Mechanisms of Change

Many child treatment researchers describe the different components of their treatment program, but few test the relative impact of the various components and even fewer provide tests of possible mediators of change. This is true despite the fact that some treatment models are quite clear about *hypothesized* mediators. Tests of these hypotheses may lead to some significant surprises. Consider, for example, the hypothesis that CBT works by changing cognitions, which in turn lead to changes in behavior. In their thoughtful meta-analysis of CBT treatment outcome studies, Durlak et al. (1991) found that changes in cognition were not significantly correlated with changes in target behavior. Such sobering findings should remind us that even when our treatments succeed, they may do so for reasons we do not yet understand. Identifying actual mediators of change may help us eliminate unnecessary elements of a treatment, strengthen effective elements, and thus enhance efficiency and impact.

Testing Varied Approaches to Treatment Delivery

Extant research on child treatment involves a rather limited range of approaches. The most common form of treatment involves group administration of a series of weekly sessions in a university laboratory clinic or school room, followed by posttreatment outcome assessment, and (in about 33% of the studies) some follow-up assessment. Group administration has several advantages (e.g., efficiency, opportunity for peer interaction), but recent success with individually administered treatment (e.g., Kendall, 1994) suggests that there may be advantages to procedures that permit individual tailoring of treatment components to fit specific child characteristics. Most child treatments currently focus rather exclusively on the children, but the recent

[5] A rationale sometimes offered for child treatment is that it is important to treat problems early in life so that they do not grow worse later. If this notion is to be properly tested, we may need to operationally define "later" in ways that go beyond 6 months.

success of efforts to involve families (e.g., Barrett et al., 1996) suggests a need to rethink. Conducting treatment sessions in a lab clinic or schoolroom also has some obvious advantages, but the recent success of treatments that take the therapist into the child's home and community (see, e.g., Henggeller, Schoenwald, & Pickrel, 1995) highlights the potential of a very different model of "delivery." Models that involve periodic posttreatment "check-up assessments," paired with booster sessions when the check-ups reveal slippage, may also bear scrutiny (see Kazdin & Weisz, 1997). The success of our treatments thus far may have been limited by overly constrained models of treatment delivery.

Building Treatments on a Broader Range of Theoretical Models

Finally, as noted earlier, a major limitation of current evidence on child treatment is that most of the treatments that have been tested are behavioral (including CBT). In the Kazdin et al. (1990) survey of 223 child treatment outcome studies, more than 70% of all the studies involved behavioral or CBT treatments, but fewer than 10% involved psychoanalytic, psychodynamic, client-centered, or existential–humanistic models. Of course, these nonbehavioral models account for a great deal of treatment by practitioners in clinical settings, and some of the treatments based on these models may work well. But we will not know until we broaden the theoretical base of child treatment outcome research. Such broadening will also, quite obviously, increase the relevance of our research to clinical practice.

Concluding Comment

There is much to like about the current state of child treatment outcome research and about the work of the Task Force in summarizing and evaluating that research. There is also much that remains to be done, by researchers and by Task Force members, to maximize the yield of research in the area and to enhance its relevance and value for clinical practice and clinical training. We have already learned a good deal about how to help children and families, but our work may have just begun.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., Rev.). Washington, DC: Author.

Baer, R. A., & Nietzel, M. T. (1991). Cognitive and behavioral treatment of impulsivity in children: A meta-analytic review of the outcome literature. *Journal of Clinical Child Psychology, 20,* 400–412.

Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64,* 333–342.

Bickman, L. (1996). A continuum of care: More is not always better. *American Psychologist, 51,* 689–701.

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388–400.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Chambless, D. L., Sanderson, W. C., Shoham, V., Johnson, S. B., Pope, K. S., Crits-Cristoph, P., Baker, M., Johnson, B., Woody, S. R., Sue, S., Beutler, L., Williams, D. A., & McCurry, S. (1996). An update on empirically validated therapies. *The Clinical Psychologist, 49,* 5–18.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Durlak, J. A., Fuhrman, T., & Lampman, C. (1991). Effectiveness of cognitive-behavior therapy for maladapting children: A meta-analysis. *Psychological Bulletin, 110,* 204–214.

Dush, D. M., Hirt, M. L., & Schroeder, H. E. (1989). Self-statement modification in the treatment of child behavior disorders: A meta-analysis. *Psychological Bulletin, 106,* 97–106.

Hazelrigg, M. D., Cooper, H. M., & Borduin, C. M. (1987). Evaluating the effectiveness of family therapies: An integrative review and analysis. *Psychological Bulletin, 101,* 428–442.

Henggeler, S. W., Schoenwald, S. K., & Pickrel, S. G. (1995). Multisystemic therapy: Bridging the gap between university- and community-based treatment. *Journal of Consulting and Clinical Psychology, 63,* 709–717.

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

Kazdin, A. E., Bass, D., Ayers, W. A., & Rodgers, A. (1990). Empirical and clinical focus of child and adolescent psychotherapy research. *Journal of Consulting and Clinical Psychology, 58,* 729–740.

Kazdin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66,* 19–36.

Kendall, P. C. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62,* 100–110.

Lochman, J. E., Coie, J. D., Underwood, M. K., & Terry, R. (1993). Effectiveness of a social relations intervention program for aggressive and nonaggressive, rejected children. *Journal of Consulting and Clinical Psychology, 61,* 1053–1058.

Mann, C. (1990). Meta-analysis in the breech. *Science, 249,* 476–480.

Pelham, W. E., Carlson, C., Sams, S., Vallano, G., Dixon, J., & Hoza, B. (1993). Separate and combined effects of methylphenidate and behavior modification on boys with Attention Deficit Hyperactivity Disorder in the classroom. *Journal of Consulting and Clinical Psychology, 61,* 506–515.

Prout, H. T., & DeMartino, R. A. (1986). A meta-analysis of school-based studies of psychotherapy. *Journal of School Psychology, 24,* 285–292.

Russell, R. L., Greenwald, S., & Shirk, S. R. (1991). Language change in child psychotherapy: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 59,* 916–919.

Saile, H., Burgmeier, R., & Schmidt, L. R. (1988). A meta-analysis of studies on psychological preparation of children facing medical procedures. *Psychology and Health, 2,* 107–132.

Shadish, W. R., Montgomery, L. M., Wilson, P., Wilson, M. R., Bright, I., & Okwumabua, T. (1993). Effects of family and marital psychotherapies: A meta-analysis. *Journal of Consulting and Clinical Psychology, 61,* 992–1002.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92,* 581–604.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Stark, K. D., & Kendall, P. C. (1995). *Treating depressed children: Therapist manual for "ACTION."* Ardmore, PA: Workbook Publishing.

Tarnowski, K. J., Rosen, L. A., McGrath, M. L., & Drabman, R. S. (1987). A modified habit reversal procedure in a recalcitrant case of trichotillomania. *Journal of Behavior Therapy and Experimental Psychiatry, 18,* 157–163.

Task Force on Promotion and Dissemination of Psychological Procedures, Division of Clinical Psychology, American Psychological Association. (1995). Training in and dissemination of empirically-validated psychological treatments: Report and recommendations. *The Clinical Psychologist, 48,* 3–23.

Weiss, B., & Weisz, J. R. (1990). The impact of methodological factors on child psychotherapy outcome research: A meta-analysis for researchers. *Journal of Abnormal Child Psychology, 18,* 639–670.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Kauneckis, D. (1995). Child and adolescent psychotherapy outcomes in experiments versus clinics: Why the disparity? *Journal of Abnormal Child Psychology, 23,* 83–106.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 688–701.

Weisz, J. R., Eastman, K. L., Donenberg, G. R., Granger, D. A., Han, S., Yeh, M., Thurber, C. A., Huey, S., Weersing, V. R., McCarty, C., & Valeri, S. (1998). [Studying clinic-based child mental health care]. Unpublished raw data.

Weisz, J. R., Han, S. S., & Valeri, S. M. (1997). More of what? Issues raised by Fort Bragg. *American Psychologist, 52,* 541–545.

Weisz, J. R., Huey, S. J., & Weersing, V. R. (1998). Psychotherapy outcome research with children and adolescents: The state of the art. In T. H. Ollendick & R. J. Prinz (Eds.), *Advances in clinical child psychology* (Vol. 20, pp. 49–91). New York: Plenum.

Weisz, J. R., Valeri, S. M., McCarty, C. A., & Moore, P. S. (in press). Interventions for child and adolescent depression: Features, effects, and future directions. In C. A. Essau & F. Petermann (Eds.), *Depressive disorders in children and adolescents.* Northvale, NJ: Harwood.

Weisz, J. R., Walter, B. R., Weiss, B., Fernandez, G. A., & Mikow, V. A. (1990). Arrests among emotionally disturbed violent and assaultive individuals following minimal versus lengthy intervention through North Carolina's Willie M. Program. *Journal of Consulting and Clinical Psychology, 58,* 720–728.

Weisz, J. R., & Weiss, B. (1989). Assessing the effects of clinic-based psychotherapy with children and adolescents. *Journal of Consulting and Clinical Psychology, 57,* 741–746.

Weisz, J. R., & Weiss, B. (1993). *Effects of psychotherapy with children and adolescents.* Newbury Park, CA: Sage.

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542–549.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117,* 450–468.

Received March 10, 1997
Revision received December 8, 1997
Accepted December 10, 1997