Across 5 decades, hundreds of randomized trials have tested psychological therapies for youth internalizing (anxiety, depression) and externalizing (misconduct, attention deficit and hyperactivity disorder) disorders and problems. Since the last broad-based youth meta-analysis in 1995, the number of trials has almost tripled and data-analytic methods have been refined. We applied these methods to the expanded study pool (447 studies; 30,431 youths), synthesizing 50 years of findings and identifying implications for research and practice. We assessed overall effect size (ES) and moderator effects using multilevel modeling to address ES dependency that is common, but typically not modeled, in meta-analyses. Mean post-treatment ES was 0.46; the probability that a youth in the treatment condition would fare better than a youth in the control condition was 63%. Effects varied according to multiple moderators, including the problem targeted in treatment: Mean ES at posttreatment was strongest for anxiety (0.61), weakest for depression (0.29), and nonsignificant for multiprob-
Mental health problems are both prevalent and disabling in children and adolescents (herein “youths”). At any one time, about one in six will meet criteria for a disorder, and at least one in three will have a disorder by age 16 (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003). A recent report in The Lancet (Gore et al., 2011) ranked neuropsychiatric disorders as the most prominent cause of the global burden of disease in young people, expressed in “years lost because of disability” (p. 2093). For many of these conditions, psychological therapy has been identified as the primary resource, highlighted as a path to promoting and protecting youth mental health, and advocated in government policy documents (U.S. Department of Health & Human Services, 2003; U.S. Public Health Service, 2000; Weisz, Sandler, Durlak, & Anton, 2005). Indeed, psychological therapy is often recommended as the first-line treatment of choice for young people, even within the practice guidelines of medical disciplines (e.g., Birmaher et al., 2007; A. H. Cheung et al., 2007; H. Steiner, 1997). But how effective is youth therapy, and what factors are associated with its level of effectiveness? Although these questions have been addressed in hundreds of randomized controlled trials (RCTs), spanning five decades, most of those trials have not yet been brought together within a single meta-analysis. A more complete and current synthesis of the evidence will clarify what is known about the impact of youth psychological therapy.

There have been previous meta-analyses of youth psychological therapy research, including four broad-based meta-analyses, encompassing a range of treatments for multiple disorders and problems. These four—Casey and Berman (1985), Kazdin, Bass, Ayers, and Rodgers (1990), Weisz, Weiss, Aliche, and Klotz (1987), and Weisz, Weiss, Han, Granger, and Morton (1995)—included 75 to 150 studies each and reported mean effect sizes (ESs) ranging from 0.54 to 0.88, medium to large effects by Cohen’s (1988) standards. Other meta-analyses used smaller study pools to investigate more specialized clinical or theoretical questions, such as how youth therapies fare with a particular disorder or problem (e.g., S. Reynolds, Wilson, Austin, & Hooper, 2012; Sonuga-Barke et al., 2013; Weisz, McCarty, & Valeri, 2006), how well a particular type of youth therapy works and what variables moderate its effects (e.g., Lundahl, Risser, & Lovejoy, 2006), whether different youth therapies differ in their effects (S. Miller, Wampold, & Varhely, 2008), and how effective therapies are for ethnic minority youths (S. J. Huey & Polo, 2008). Broad-based meta-analysis can be seen as a useful complement to these more focused syntheses and to individual studies that provide precise tests of specific hypotheses. Because each approach has strengths and limitations, the combination of big-picture breadth and more narrowly focused precision offers the potential for valuable triangulation.

As stressed by many in the field (e.g., Atkins, Fink, Slutsy, Agency for Healthcare Research and Quality, & North American Evidence-Based Practice Centers, 2005; Eccles, Freemantle, & Mason, 2001), including the Cochrane Collaboration (Higgins, Green, & Scholten, 2011), meta-analyses and systematic reviews are most useful if periodically updated. Failure to publish updates may leave researchers, clinicians, and policymakers acting on out-of-date or misleading evidence. Cumulative knowledge is best reflected when current evidence synthesis incorporates and builds upon previous syntheses that are already in the evidence base. The overlap helps to reduce gaps and ensures a more complete picture of the state of knowledge, particularly when new candidate moderators (e.g., study variables evaluated for their impact on standardized mean differences between treatment and control conditions) are tested and refined data analytic methods are employed (see also Moher et al., 2007; Shojania et al., 2007).

The need for an updated synthesis is especially clear in the case of research on youth psychological therapy. Two decades have passed since the last broad-based synthesis of the youth therapy evidence base encompassing multiple problems and multiple treatments. During that time, the number of peer-reviewed youth RCTs has almost tripled, such that any conclusions derived from previous broad-based meta-analyses do not reflect even half of the evidence now available. Moreover, the previous findings were derived from meta-analytic methods of the time that could not directly address within-trial dependency among ESs, which is pervasive in therapy research. It is now possible to use a multilevel approach to directly address, and even model, the...
dependency. The large, untapped evidence base and the recently refined methods make this an excellent time to examine what is known about youth therapy impact and to consider what future directions the findings suggest for research and practice.

We conducted a meta-analysis of the greatly expanded evidence base (447 studies vs. 150 studies in the most recent broad-based youth meta-analysis; Weisz et al., 1995) to learn what 50 years of research has shown. We included RCTs across multiple domains, encompassing internalizing and externalizing problems, to address questions that cannot be answered with a focus on only one disorder, one problem category, one therapy method, or one type of control condition. In its breadth of coverage, this synthesis complements more narrowly focused syntheses found in separate publications (e.g., S. J. Huey & Polo, 2008; Lundahl et al., 2006; S. Miller et al., 2008; S. Reynolds et al., 2012; Weisz et al., 2013). Our broad synthesis of studies was used to address four primary research questions:

1. What is the overall effect of youth psychological therapy? Previous broad-based meta-analyses (Casey & Berman, 1985; Kazdin et al., 1990; Weisz et al., 1987, 1995) generated mean ESs that may no longer be representative of the evidence base. Moreover, their accuracy may have been undermined by less-than-ideal methods for addressing ES dependency (Lipsey & Wilson, 2000). Most RCTs generate numerous ESs, in part because they include multiple outcome measures and multiple informants. In most previous meta-analyses, the resulting dependency—if addressed—has been dealt with by either choosing only one ES from many or averaging across ESs within studies. Such approaches lose information, rule out critical analyses of within-study factors (e.g., informant differences), artificially reduce the variance between ESs, and—importantly—risk inaccurate estimation of study ESs (Becker, 2000; S. F. Cheung & Chan, 2004, 2008). We used a three-level random effects model that permits incorporating all relevant ESs from each study while also accounting for dependency among these ESs (Van den Noortgate, Lopez-Lopez, Marin-Martinez, & Sanchez-Meca, 2013). Applying this approach to the 50-year RCT database helped to maximize the accuracy of the overall estimate of treatment impact and its variability between and within studies.

2. Does therapy impact differ by target problem? Previous broad-based meta-analyses have not addressed this question well because they have not reported effects for separate targeted problem areas, or they have merged distinct categories—for example, combining anxiety and depression within an “overcontrolled-internalizing” category (Weisz et al., 1995). We compared the findings of studies targeting anxiety, depression, attention-deficit hyperactivity disorder (ADHD), and conduct problems—the four most common youth problem domains (Weisz, 2004). Previous youth therapy reviews noted positive effects within problem areas (e.g., David-Ferdon & Kaslow, 2008; Eyberg, Nelson, & Boggs, 2008; Kendall, Settipani, & Cummings, 2012), but the present meta-analysis is the first comparison of effects across these four target problem domains drawing on a large representative pool of studies. As a complement, we added a fifth category: treatments targeting multiple problems concurrently. Clinically referred youths show high rates of comorbidity (Angold, Costello, & Erkanli, 1999; Costello et al., 2003). Most youth therapy studies, although their samples may have had high comorbidity, have focused on treatment of a particular target problem (e.g., anxiety). However, some investigators have dealt with comorbidity and co-occurring problems by attempting to treat multiple problem domains concurrently within the same treatment. Indeed, some have stressed that treatments focused on single disorders or problem areas may need to be replaced or complemented by treatments that encompass multiple different problems within the same protocol (see, e.g., Barlow & Carl, 2010). To find out how the multiproblem approaches tried thus far
have fared, we assessed the effects of the multi-problem treatments tested across five decades.

3. Does therapy impact differ according to type of therapy? There is some meta-analytic support for the view that different youth therapies differ in their effects; for example, analyses by Weiss and Weisz (1995) and Weisz et al. (1995) showed larger effects for behavioral (including cognitive-behavioral) than nonbehavioral therapies. But other evidence supports the view “that when psychological therapies intended to be therapeutic are compared, the true differences among all such treatments are 0” (Wampold et al., 1997, p. 203). This perspective has been named for the dodo bird’s conclusion in Alice in Wonderland that “everybody has won, and all must have prizes” (Carroll, 1865/1962). Wampold et al. (1997) found support for the dodo bird view in a meta-analysis that included adult studies, and S. Miller et al. (2008) found mixed support in a meta-analysis of 23 youth therapy studies directly pitting active treatments against one another. Such treatment versus treatment studies can provide evidence directly relevant to the dodo bird hypothesis. However, the number of studies comparing any two treatment approaches tends to be small, limiting power to detect differences; and Miller et al. have found that investigator allegiance in such studies may bias findings and complicate their interpretation. Thus, it is useful to complement syntheses of treatment versus treatment studies with meta-analyses that compare treatment effects in treatment versus control studies. Such studies are numerous and less likely to be compromised by differential allegiance (because investigator bias favoring a control condition is unlikely). We examined the dodo bird hypothesis using youth treatment versus control condition RCTs and tested whether the effect of treatment type might depend on other factors, such as the source of outcome data.

4. Does therapy impact depend on the control condition employed? Although a treatment’s ES is influenced by what that treatment is compared with, control condition effects have not been featured in many meta-analyses. One meta-analysis (Kazdin et al., 1990) found larger effects with passive versus active control conditions, and others (Spielmans, Gatlin, & McFall, 2010; Weisz et al., 2013) have suggested that “usual care” is an important externally valid control condition to evaluate. We compared effects for passive control conditions and four kinds of active controls, including usual care.

We tested three primary candidate moderators—targeted problem, therapy type, and control condition—and interactions between each pair of moderators. As a secondary focus, we tested eight additional candidate moderators highlighted in prior research (Casey & Berman, 1985; De Los Reyes et al., 2015; Kazdin et al., 1990; Weisz & Kazdin, 2010; Weisz et al., 1987, 1995, 2005): study year, study location, how study participants were identified, ethnicity, gender, developmental period, whether diagnosis was required, and which informant reported on outcome. We tested the effects of these eight variables and the relationship of each variable to the three primary candidate moderators.

Method

Data Sources and Study Selection

The literature search included peer-reviewed RCTs testing psychological therapies for youth psychopathology encompassing four broad domains that account for most mental health referrals (Weisz, 2004; Weisz & Kazdin, 2010)—depression (e.g., Mufson, Weissman, Moreau, & Garfinkel, 1999), anxiety (e.g., Kendall et al., 1997; the anxiety category included obsessive-compulsive disorder and posttraumatic stress disorder), conduct problems (e.g., Scott et al., 2010), and ADHD (e.g., MTA Cooperative Group, 1999a). Because we focused on peer-reviewed RCTs, we did not include unpublished dissertations. We searched multiple sources. These included PsycINFO and PubMed from Jan-
of psychopathology are common in the youth treatment literature, for several reasons: (a) both definitions of psychopathology as either meeting criteria for a disorder or showing elevated symptoms (e.g., scoring in the clinical range on standardized measures of psychopathology), for several reasons: (a) both definitions of psychopathology are common in the youth treatment outcome literature (Weisz, 2004; Weisz & Kazdin, 2010); (b) youths with elevated behavioral or emotional symptoms experience serious impairment (Angold, Costello, Farmer, Burns, & Erkanli, 1999; Costello, Angold, & Keeler, 1999; Silverman & Hinshaw, 2008; Weisz, 2004); (c) such youths are often referred for mental health services (A. L. Jensen & Weisz, 2002; Weisz, Ugueto, Cheron, & Herren, 2013); and (d) diagnostic categories and their definitions within the DSM-IV-TR system have varied markedly across the five decades encompassed in this meta-analysis. Figure 1 shows the study search and identification flowchart.

Data Extraction, Coding, and Processing

Studies were coded for study and sample characteristics, treatment procedures, study quality indicators, and multiple candidate outcome moderators (see Table 1). Clarity of authors’ reporting varied from study to study, making it important to assess intercoder agreement. To assess agreement, eight coders independently coded 20 to 30 randomly selected studies each; the most experienced RCT coder (a RCT researcher with a doctoral degree in clinical psychology) served as the master coder against which responses of the other coders (graduate students and postdoctoral fellows in clinical psychology) were compared. Intercoder agreement is reported in the next paragraph. When coder disagreements arose, these were resolved by discussion, together with review of the relevant portions of articles being coded.

We coded the year and location of each study within versus outside North America (k = .80). The participant engagement variable (k = .66) included codes for samples that were either (a) recruited for the study, (b) clinically referred/treatment-seeking (i.e., at outpatient clinics, impatient clinics, or school-based mental health services), or (c) receiving treatment on a nonvoluntary basis because of court mandate or incarceration. We coded the percent of Caucasian participants (intraclass correlation coefficient [ICC] = .87), percent of each gender (ICC = .95), and mean participant age (ICC = .99), and then dichotomized these variables into majority (>50%) Caucasian versus minority, majority (>50%) male versus female, and majority children (mean <12 years) versus adolescents for moderator analyses. We also coded whether study participants were required to have a diagnosis in the targeted problem area or not (k = .79), and type of targeted problem—ADHD, conduct, anxiety, depression, or multiple problems (k = .89). The informant variable (k = .87) included youths, parents, and teachers (other informants [e.g., therapists, siblings]

1 Our search accessed the earliest available version of each relevant article; in recent years, this was often an online advance version. In the References for the present article, we provide publication year and details for the hard copy print version of those articles that have since been published.
were excluded from moderator analyses because of low \( n \). In addition, we coded the number of weeks from baseline to each assessment point (ICC = 0.98). Finally, we coded treatment and control conditions (\( k = 0.83 \)).

Treatment type codes were collapsed into four broad orientations: (a) youth-focused behavioral interventions (i.e., CBT, modeling, psychoeducation, operant or respondent conditioning, social skills training, biofeedback, behavioral activation, or a combination of these; the treatments could be individually or group-administered), (b) youth-focused nonbehavioral interventions (i.e., client-centered, psychodynamic, gestalt therapies, or a combination of these), (c) caregiver and family-focused behavioral treatment (i.e., behavioral parent training, behavioral parent–youth/family interventions, such as functional family therapy), (d) caregiver and family-focused nonbehavioral treatment (i.e., nonbehavioral parent, parent–youth, or family interventions, such as attachment-based family therapy) and multisystem treatment (i.e., multisystemic therapy, Treatment Foster Care Oregon—previously known as Multidimensional Treatment Foster Care, or other multisystem intervention). Moderator analyses involving treatment type excluded treatments that combined interventions across the four broad categories and treatments that appeared rarely and did not fit any of the broad categories (e.g., Interpersonal Therapy for Adolescents). However, studies of all treatment types were included in the analyses that did not test the treatment type variable (e.g., in analyses of ES for the different targeted problems). Control conditions were no-treatment/waitlist, therapy placebo, pill placebo, and usual clinical care, in which therapists used what-

---

**Dikla Eckshtain**

---

**Figure 1.** Flowchart showing study retrieval, review, exclusion, and inclusion.
ever treatments they used in their usual practice and the treatments could not be coded as a single treatment type.

In obtaining and processing the data and reporting findings, our procedures were consistent with PRISMA guidelines (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009), derived from the earlier QUORUM statement on meta-analysis reports (Moher et al., 1999), with a few practical exceptions. For example, because of our inclusion of 447 studies, it was not feasible to include a table providing details of every separate study.

**Effect Size Calculation**

ESs were represented as Cohen’s $d$ (Cohen, 1988), reflecting the standardized mean difference between treatment and control conditions. Our standard ES calculations used Cohen’s $d$. For each study, we calculated a single ES. The ESs were weighted by the total sample size for each study, which allowed us to compare studies of different sample sizes. We used a random-effects model to account for the variability among studies. We then used a mixed-effects model to analyze the data, which allowed us to account for the nested structure of the data (i.e., ESs within studies). The model included fixed effects for moderators and random effects for the residual variation.

Table 1

<table>
<thead>
<tr>
<th>Moderator</th>
<th>$n$ of studies</th>
<th>$n$ of ESs</th>
<th>Subgroup analysis</th>
<th>Moderator test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ES (g)</td>
<td>95% CI</td>
<td>Test statistic</td>
</tr>
<tr>
<td>Study-level moderators (third level)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study year</td>
<td>443</td>
<td>5,139</td>
<td></td>
<td>$t(5137) = -2.24$</td>
</tr>
<tr>
<td>Study location</td>
<td>443</td>
<td>5,139</td>
<td></td>
<td>$t(5137) = .64$</td>
</tr>
<tr>
<td>North America</td>
<td>303</td>
<td>3,680</td>
<td>.42***</td>
<td>.36 .48</td>
</tr>
<tr>
<td>Outside North America</td>
<td>140</td>
<td>1,459</td>
<td>.54***</td>
<td>.45 .62</td>
</tr>
<tr>
<td>Participant engagement</td>
<td>406</td>
<td>4,688</td>
<td></td>
<td>$F(2, 4485) = 2.00$</td>
</tr>
<tr>
<td>Recruited</td>
<td>276</td>
<td>3,225</td>
<td>.50***</td>
<td>.44 .56</td>
</tr>
<tr>
<td>Referred</td>
<td>92</td>
<td>1,087</td>
<td>.41***</td>
<td>.30 .52</td>
</tr>
<tr>
<td>Nonvoluntary</td>
<td>38</td>
<td>376</td>
<td>.35***</td>
<td>.19 .52</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>224</td>
<td>2,763</td>
<td>.46***</td>
<td>.37 .55</td>
</tr>
<tr>
<td>Caucasian sample (≥50% Caucasian)</td>
<td>156</td>
<td>2,101</td>
<td>.49***</td>
<td>.36 .62</td>
</tr>
<tr>
<td>Non-Caucasian sample (&lt;50% Caucasian)</td>
<td>68</td>
<td>662</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>384</td>
<td>4,696</td>
<td></td>
<td>$t(4694) = -.39$</td>
</tr>
<tr>
<td>Majority male (&gt;50% male)</td>
<td>269</td>
<td>3,557</td>
<td>.47***</td>
<td>.41 .53</td>
</tr>
<tr>
<td>Majority female (&gt;50% female)</td>
<td>115</td>
<td>1,139</td>
<td>.44***</td>
<td>.35 .54</td>
</tr>
<tr>
<td>Developmental period</td>
<td>441</td>
<td>5,132</td>
<td></td>
<td>$t(5130) = -1.55$</td>
</tr>
<tr>
<td>Childhood (mean age &lt;12 years)</td>
<td>286</td>
<td>3,622</td>
<td>.49***</td>
<td>.42 .55</td>
</tr>
<tr>
<td>Adolescence (mean age ≥12 years)</td>
<td>155</td>
<td>1,510</td>
<td>.40***</td>
<td>.32 .49</td>
</tr>
<tr>
<td>Diagnosis requirement</td>
<td>193</td>
<td>2,768</td>
<td></td>
<td>$t(2766) = -0.56$</td>
</tr>
<tr>
<td>Required of all participants</td>
<td>149</td>
<td>2,272</td>
<td>.50***</td>
<td>.41 .59</td>
</tr>
<tr>
<td>Not required</td>
<td>44</td>
<td>496</td>
<td>.45***</td>
<td>.27 .62</td>
</tr>
<tr>
<td>Targeted problem</td>
<td>443</td>
<td>5,139</td>
<td></td>
<td>$F(4, 5134) = 6.62$</td>
</tr>
<tr>
<td>ADHD$^b$</td>
<td>82</td>
<td>1,161</td>
<td>.34***</td>
<td>.23 .45</td>
</tr>
<tr>
<td>Conduct$^{b,c,d}$</td>
<td>158</td>
<td>1,952</td>
<td>.46***</td>
<td>.38 .54</td>
</tr>
<tr>
<td>Anxiety$^{b,c,d}$</td>
<td>143</td>
<td>1,493</td>
<td>.61***</td>
<td>.53 .70</td>
</tr>
<tr>
<td>Depression$^{c,e}$</td>
<td>47</td>
<td>443</td>
<td>.29***</td>
<td>.14 .43</td>
</tr>
<tr>
<td>Multiple problems$^{d,f}$</td>
<td>13</td>
<td>90</td>
<td>.15</td>
<td>-.14 .43</td>
</tr>
</tbody>
</table>

ES-level moderators (second level)

| Informant | 379          | 3,399     |                   | $F(2, 3396) = 23.27$ | <.001           |
| Youth$^{b,c}$ | 239  | 1,342     | .43***            | .37 .49       |               |
| Parent$^{c}$ | 209  | 1,598     | .48***            | .42 .54       |               |
| Teacher$^{b,c}$ | 108  | 459       | .27***            | .20 .35       |               |
| Treatment type$^c$ | 347  | 3,669     |                   | $F(3, 3694) = 1.46$ | .123           |
| Youth-focused behavioral | 239  | 2,216     | .48***            | .41 .54       |               |
| Youth-focused nonbehavioral | 29   | 148       | .33***            | .17 .50       |               |
| Caregiver/Family-focused behavioral | 78  | 1,092     | .44***            | .35 .53       |               |
| Caregiver/Family-focused nonbehavioral | 6    | 43        | .45*              | .04 .87       |               |
| Multisystem | 15    | 200       | .25               | -.01 .51      |               |
| Control condition | 443  | 5,139     |                   | $F(4, 5134) = 3.44$ | .008           |
| No treatment/waitlist$^a,b$ | 234  | 2,776     | .53***            | .47 .59       |               |
| Psychotherapy placebo$^a$ | 113  | 1,173     | .41***            | .33 .50       |               |
| Pill placebo | 5     | 43        | .37               | -.08 .83      |               |
| Case management | 34    | 322       | .39***            | .21 .57       |               |
| Usual care treatment$^b$ | 59    | 597       | .30***            | .16 .43       |               |

Note. Some moderators were missing for certain studies. Each study can contribute multiple ESs; thus, study sample size across subgroups can exceed the total study sample size for the ES-level moderators. Within each moderator having more than two subgroups, identical letter superscripts indicate significant ($p < .05$) pairwise comparisons between subgroups. ES = effect size; $g$ = Hedges’ $g$; CI = confidence interval; ADHD = attention-deficit hyperactivity disorder.

$^a$ The behavioral categories included behavioral and cognitive-behavioral therapies. $^{***} p < .001$. 

Reference:


Data Synthesis

Multilevel meta-analytic approach. Most studies (96%) yielded multiple ESs, so the assumption of independence of ESs that underlies traditional meta-analytic approaches was violated. As noted earlier, dependency among ESs has commonly been dealt with by either choosing only one ES from many or averaging across ESs within studies in order to obtain a single ES per study, and notable shortcomings have been associated with such approaches (see, e.g., Becker, 2000; S. F. Cheung & Chan, 2004, 2008). We addressed the dependency issue by using a multilevel approach that permitted us to include all ESs derived from measures of the targeted problems, in nonaggregated form for each study. The three-level model encompassed the sampling variation for each ES (Level 1), within-study variation (Level 2), and between-study variation (Level 3). The model consists of three regression equations, one for each level:

\[ d_{jk} = \beta_{0jk} + r_{jk} \quad \text{with} \quad r_{jk} \sim N(0, \sigma^2_{rjk}) \]  
\[ \beta_{0jk} = \theta_{00k} + u_{0jk} \quad \text{with} \quad u_{0jk} \sim N(0, \sigma^2_u) \]  
\[ \theta_{00k} = \gamma_{000} + \gamma_{00k} \quad \text{with} \quad \gamma_{00k} \sim N(0, \sigma^2_\gamma) \]  

Equation 1 indicates that the \( j \)th observed ES from study \( k \) equals its population value, plus a random deviation, which is assumed to be normally distributed. This residual variance is estimated before performing the meta-analysis, which comprised the mean observed sampling variance of the standardized mean difference (\( d \)) in this study. Equation 2 states that the population values comprise a study mean and random deviation from this mean, which is assumed to be normally distributed. In Equation 3, study mean effects are assumed to vary randomly around an overall mean. This extension of the commonly used random-effects meta-analytic model provided an overall estimate of the difference between treatment and control conditions.

We subsequently fitted a three-level mixed effects model to identify moderators that might explain variation in ESs between and within studies by adding study (Level 3) or outcome (Level 2) characteristics as fixed predictors. Examining treatment outcome in this meta-analysis of RCTs involved synthesizing standardized mean differences between treatment versus control condition and thus entailed comparison of conditions within each study. In this context, “treatment outcome” is in fact the relation between study condition and outcome. If this relationship varies according to a study or outcome characteristic, then study condition interacts with that characteristic, which is thus called a moderator (see Hedges & Pigott, 2004; Kraemer, Wilson, Fairburn, & Agras, 2002). In the present meta-analysis, we considered the three primary (i.e., target problem, type of treatment, control condition) and eight secondary candidate moderators (i.e., study year, study location, participant engagement, ethnicity, gender, developmental period, diagnosis requirement, informant). Simple moderator analyses were only conducted if each category contained at least five studies (parameter estimates are poor when number of studies is very small). Because including multiple categorical moderators may inflate Type II error rates (Raudenbush & Bryk, 2002; Van den Noortgate et al., 2013), separate three-level mixed models were fitted for each candidate moderator variable at each time point (i.e., posttreatment and follow-up). Although moderators are the keys to explaining ES differences, they may also be interrelated, complicating the interpretation of simple moderator effects. To examine possible relations among moderators, we examined interactions among moderators to test whether moderator effects differed as a function of each of the three main moderators.
(i.e., target problem, type of treatment, control condition). We used a parsimonious modeling approach, testing these complex moderator effects one at a time, and applying a Bonferroni correction ($p < .005$) to the tests.

The parameters estimated in a multilevel meta-analysis are the regression coefficients of the highest level equations and the variances at the second and third level. Fixed model parameters are tested using a Wald test, which compares the difference in parameter estimate and the hypothesized population value, divided by the standard error, with a $t$-distribution. For categorical variables with more than two categories, the omnibus test follows an $F$-distribution and pairwise comparisons were used to test which subgroup mean ESs were significantly different. Similar to standard random effects meta-analysis, the between-study heterogeneity is reflected by the between-study (Level 3) variance, but the three-level model also yields an estimate of the within-study variance (Level 2). Likelihood ratio tests, comparing the deviance scores of the full model and models excluding the variance parameter at a certain level, were used to test the variation in ESs between and within studies. Parameters were estimated using the restricted maximum likelihood procedure implemented in SAS PROC HP-MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Observed ESs were weighted by the inverse of the sampling variance, and the residual degrees of freedom was used to compute the denominator degrees of freedom for the fixed effects.

Publication bias. The risk that studies with null or negative findings were less likely to be published (Begg, 1994; Lipsey & Wilson, 2000; Rosenthal, 1979) was addressed in three ways. First, we used a funnel plot (Torgerson, 2006), with standard error plotted on the vertical axis as a function of ES on the horizontal axis. In the absence of publication bias, the plot resembles an inverted funnel with studies distributed symmetrically around the mean ES (Begg, 1994). Egger’s weighted regression test (Egger, Davey Smith, Schneider, & Minder, 1999) was used to assess plot asymmetry. Second, the trim-and-fill method was used to correct funnel plot asymmetry arising from publication bias (Duval & Tweedie, 2000). Third, we computed a fail-safe $N$, the number of additional “zero-effect” studies needed to increase the $p$ value for the meta-analysis to above .05 (Rosenthal, 1979). These analyses were conducted using the R software package metafor (Viechtbauer, 2010).

Study quality. We coded each study using methodological quality variables employed in previous reviews (Higgins & Green, 2008; Kocsis et al., 2010; Moher et al., 1998; Weiss & Weisz, 1990) that were reported with sufficient frequency and clarity to be applied to the majority of studies and coded reliably. The variables that met these criteria were (a) subject blindness to assessment (that is, whether the person who was the focus of assessment was aware of being assessed and able to influence assessment ratings or scores; mean $k = .60$); (b) participant attrition (that is, the percentage of participants at randomization who were available for ES computation; mean ICC for group sample sizes $= 0.98$); (c) measurement objectivity (that is, subjective [e.g., youth report and parent report] versus objective [for example, task performance and behavior coding measures were coded as objective and self-report as not objective]; mean $k = .86$); and, for the treatment conditions, presence of pretherapy training (mean $k = .74$), adherence/fidelity checks (mean $k = .77$), and treatment manual or structured guide (mean $k = .60$). The impact of these six variables on treatment effects was assessed using three-level mixed models with Bonferroni adjustment ($p < .008$) to address the risk of chance findings.

Results

Study Pool

Four hundred forty-seven RCTs met inclusion criteria, generating 6,941 ESs (the studies are identified in the References with asterisks). At posttherapy assessment, the studies, spanning 1963 to 2013, included 30,431 participants (mean study $n = 68.69$; mean group $n = 27.63$). Mean age was 9.84 years ($SD = 3.85$); mean percent male was 63.75 ($SD = 27.63$). Mean age was 9.84 years ($SD = 3.85$); mean percent male was 63.75 ($SD = 24.00$). Some 443 studies assessed outcomes at posttherapy. The treatment protocols studied specified a mean of 16.54 sessions ($SD = 14.06$; number of sessions was not significantly related to ES, $r(4,347) = -1.44, p = .149$). Time in treatment averaged 15.81 weeks ($SD = 14.72$; number of weeks in treatment was significantly,
negatively related to ES, \(t(4,664) = -2.20, p = .028\). Some 140 studies included follow-up assessment, averaging 44.26 weeks after the end of treatment (SD = 6.28).

Posttreatment Results

The 443 studies reporting outcomes at posttreatment produced 5,139 dependent ESs. Mean posttreatment ES was 0.46 (95% confidence interval [CI] [0.41, 0.51]), \(t(5138) = 18.11, p < .001\). Between-study variance was significant \((\sigma^2 = 0.223), \chi^2(1) = 1,792.4, p < .001\), as was within-

study variance \((\sigma^2 = 0.143), \chi^2(1) = 3,437.4, p < .001\), with a mean observed sampling (residual) variance of 0.143. Of the total variance, 45% was attributable to between-study differences and 28% to within-study differences.

As shown in Table 1, there were significant effects for two of the three primary moderators. Pairwise comparisons showed that effects were larger for anxiety than ADHD, conduct, depression, or multiple problems, and larger for conduct than depression or multiple problems (see Figure 2); and larger with no-treatment or waitlist controls than with therapy placebo or usual care. Moderator analyses also indicated that effects were larger in studies done outside, than within, North America, and larger for outcomes reported by youths and parents than by teachers.

We found three Moderator \(\times\) Moderator interactions, each involving the informant moderator. We dismantled these three interactions from both possible directions, beginning by examining informant effects at each level of the other three moderators (Table 2). First, the informant effect differed significantly by target problem. Significant differences between informants were found for depression, in which parent reports showed smaller effects than both teacher and youth reports. In fact, by teacher report, treatments for depression were less effective than control conditions. For ADHD and conduct, parent reports yielded larger effects than youth and teacher reports. Second, the informant effect differed by control condition; informant was significant only for no-treatment/waitlist controls, with parents and youths reporting larger effects than teachers, and for usual care controls, with teacher reports showing smaller effects than youth and parent reports. Third, the informant effect differed by treatment type, with significant

![Figure 2](https://example.com/figure2.png)

Figure 2. Mean effect sizes at posttreatment and follow-up by target problem. Dashed horizontal lines show mean effects for the full sample of studies reporting posttreatment assessments, and for the full sample of studies reporting follow-up assessments. Error bars represent standard error. Effect sizes at posttreatment and follow-up for anxiety, conduct, ADHD, and depression have error bars that do not cross the x-axis, indicating that all these effect sizes are significantly greater than zero. See the online article for the color version of this figure.
Informant effects for each targeted problem and treatment type within each type of informant. Both tables show that, of the different treatment types, only youth-focused behavioral therapies (including CBT) produced significant treatment effects across all three informants (see Figure 3).

To examine variance explained by study and outcome characteristics, we fitted a three-level mixed model including all significant moderator effects and all significant interactions among moderators. Between-study variance was reduced by 4%, and within-study variance by 12%.

Follow-Up Results

The 140 studies reporting follow-up assessments generated 1,802 dependent ESs. Mean follow-up ES was 0.36 (95% CI [0.27, 0.44]), t(1801) = 8.30, p < .001. This mean was not significantly different from the mean posttreatment ES across all studies, t(6939) = −0.003, p = .998, nor for the subset of studies (n = 136) that included both posttreatment and follow-up assessments, t(3445) = −0.47, p = .635. Between-study variance was significant (τ² = 0.218), χ²(1) = 700.4, p < .001, as was within-study variance (τ² = 0.120), χ²(1) = 2040.2, p < .001, with residual (sampling) variance of 0.121. Some 48% of the total variance was attributable to between-study differences, and 26% to within-study differences.

The same moderators tested in posttreatment analyses were examined for follow-up assessments. The analyses showed effects for target problem and informant (see Table

Table 2

Interactions Among Moderators at Posttreatment: Examined by Informant Effects

<table>
<thead>
<tr>
<th>Informant effects for each target problem, treatment type, and control condition</th>
<th>Test statistic</th>
<th>p</th>
<th>Youth</th>
<th>Parent</th>
<th>Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informant effects for each targeted problem</td>
<td>F(7, 3385) = 4.40</td>
<td>&lt;.001</td>
<td>.11**</td>
<td>.35**</td>
<td>.18**</td>
</tr>
<tr>
<td>ADHD</td>
<td>F(2, 3385) = 9.15</td>
<td>&lt;.001</td>
<td>.40***</td>
<td>.51***</td>
<td>.26***</td>
</tr>
<tr>
<td>Conduct</td>
<td>F(2, 3385) = 16.14</td>
<td>&lt;.001</td>
<td>.58***</td>
<td>.64***</td>
<td>.44***</td>
</tr>
<tr>
<td>Anxiety</td>
<td>F(2, 3385) = 2.94</td>
<td>.053</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>F(2, 3385) = 9.65</td>
<td>&lt;.001</td>
<td>.32***</td>
<td>.15**</td>
<td>(−.41)**</td>
</tr>
<tr>
<td>Multiple problems</td>
<td>F(1, 3385) = 7.03</td>
<td>.311</td>
<td>.10</td>
<td>.25</td>
<td></td>
</tr>
<tr>
<td>Informant effects for each treatment type†</td>
<td>F(6, 2340) = 7.25</td>
<td>&lt;.001</td>
<td>.43***</td>
<td>.45***</td>
<td>.44***</td>
</tr>
<tr>
<td>Youth-focused behavioral</td>
<td>F(2, 2340) = 10</td>
<td>.090</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth-focused nonbehaviorous</td>
<td>F(2, 2340) = 4.87</td>
<td>.008</td>
<td>.50***</td>
<td>(.18)</td>
<td>−.05*</td>
</tr>
<tr>
<td>Caregiver/Family-focused behavioral</td>
<td>F(2, 2340) = 35.80</td>
<td>&lt;.001</td>
<td>.36***</td>
<td>.51***</td>
<td>−.06**</td>
</tr>
<tr>
<td>Caregiver/Family-focused nonbehavioral</td>
<td>F(2, 2340) = .01</td>
<td>.907</td>
<td>.55*</td>
<td>.58*</td>
<td></td>
</tr>
<tr>
<td>Multisystem</td>
<td>F(1, 2340) = 10</td>
<td>.753</td>
<td>.13</td>
<td>.11</td>
<td></td>
</tr>
<tr>
<td>Informant effects for each control condition</td>
<td>F(8, 3384) = 4.27</td>
<td>&lt;.001</td>
<td>.51***</td>
<td>.60***</td>
<td>.36***</td>
</tr>
<tr>
<td>No treatment/waitlist</td>
<td>F(2, 3384) = 20.30</td>
<td>&lt;.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy placebo</td>
<td>F(2, 3384) = 2.13</td>
<td>.119</td>
<td>.35***</td>
<td>.24***</td>
<td>.34***</td>
</tr>
<tr>
<td>Pill placebo</td>
<td>F(2, 3384) = .06</td>
<td>.939</td>
<td>(.07)</td>
<td>(.02)</td>
<td>(.17)</td>
</tr>
<tr>
<td>Case management</td>
<td>F(2, 3384) = .97</td>
<td>.380</td>
<td>.34**</td>
<td>.27**</td>
<td>(−.001)</td>
</tr>
<tr>
<td>Usual care treatment</td>
<td>F(2, 3384) = 16.89</td>
<td>&lt;.001</td>
<td>.21**</td>
<td>.32***</td>
<td>−.11**</td>
</tr>
</tbody>
</table>

Note. Bonferroni correction was used (p set at <.005) when testing two-way interactions for each clinical variable. Asterisks denote that the ES was significantly different from zero; identical letter superscripts within each row denote significant pairwise comparisons; a dash denotes zero studies in that subgroup; ES estimates in parentheses are based on fewer than five studies. ES = effect size; ADHD = attention-deficit hyperactivity disorder.
† The behavioral categories included behavioral and cognitive-behavioral therapies. *p < .05. **p < .01. ***p < .001.
4). Pairwise comparisons revealed larger effects for anxiety than for ADHD, depression, and multiple problems, and smaller effects for teacher-reported outcomes than for youth- or parent-reported outcomes. Tests for interactions among moderators were not conducted at follow-up; parameter estimates would have been poor because more than one third of the cells contained fewer than five studies.

To examine variance explained by study and outcome characteristics, we fitted a three-level mixed model including all significant moderator effects. Between-study variance was reduced by 27%, and within-study variance by 45%.

**Publication Bias and Study Quality**

To test the robustness of the estimated effects, we assessed publication bias and study quality. Absent publication bias, the plot should resemble an inverted funnel with studies distributed symmetrically around the mean ES. Egger’s weighted regression test showed our plot to be asymmetrical, $t(445) = 6.34$, $p < .001$; with increasing standard errors, more highly positive ESs were obtained. This could indicate some degree of publication bias, or an overestimate of treatment effects in smaller studies, which are generally less rigorous (Jini, Holenstein, Sterne, Bartlett, & Egger, 2002). Importantly, when the trim-and-fill procedure was applied, the adjusted ES remained unchanged, suggesting minimal impact of publication bias. In addition, the fail-safe $N$ showed that 90,654 studies with mean ES = .00 would need to be incorporated into the meta-analysis to yield a nonsignificant effect. This markedly exceeded the Rosenthal (1979) benchmark of 2,245 ($5n + 10$), suggesting that our findings are robust to the threat that excluded studies might have yielded a nonsignificant effect. In addition, none of the six study quality variables had a significa-

**Table 3**

**Interactions Among Moderators at Posttreatment: Examined by Targeted Problem, Treatment Type, and Control Condition Effects**

<table>
<thead>
<tr>
<th>Moderator effects for each informant</th>
<th>Test statistic</th>
<th>$p$</th>
<th>ADHD</th>
<th>Conduct</th>
<th>Anxiety</th>
<th>Depression</th>
<th>Multiple problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP effects by informant</td>
<td>$F(7, 3385) = 4.40$</td>
<td>&lt;.001</td>
<td>.11&lt;sup&gt;a&lt;/sup&gt;&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.40&lt;sup&gt;c&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.58&lt;sup&gt;b,d&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.32&lt;sup&gt;d&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.10&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>TP for youth</td>
<td>$F(4, 3385) = 6.91$</td>
<td>&lt;.001</td>
<td>.35&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.51&lt;sup&gt;b&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.64&lt;sup&gt;b&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.15&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.25&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>TP for parent</td>
<td>$F(4, 3385) = 6.38$</td>
<td>&lt;.001</td>
<td>.18&lt;sup&gt;c&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.26&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.44&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.40&lt;sup&gt;a&lt;/sup&gt;&lt;sup&gt;c&lt;/sup&gt;</td>
<td>---</td>
</tr>
<tr>
<td>TP for teacher</td>
<td>$F(3, 3385) = 5.16$</td>
<td>.002</td>
<td>.50&lt;sup&gt;c&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.58&lt;sup&gt;b&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.32&lt;sup&gt;d&lt;/sup&gt;&lt;sup&gt;+++&lt;/sup&gt;</td>
<td>.34&lt;sup&gt;**&lt;/sup&gt;</td>
<td>.21&lt;sup&gt;**&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

| Youth-focused behavioral            | --- | --- | --- | --- | --- | --- |
| Youth-focused nonbehavioral         | --- | --- | --- | --- | --- | --- |
| Caregiver/Family-focused behavioral | --- | --- | --- | --- | --- | --- |
| Caregiver/Family-focused nonbehavioral | --- | --- | --- | --- | --- | --- |
| Multisystem                         | --- | --- | --- | --- | --- | --- |

| TT effects by informant<sup>f</sup> | $F(5, 2340) = 8.64$ | <.001 | .43<sup>+++</sup> | .50<sup>c</sup><sup>+++</sup> | .36<sup>+++</sup> | .55<sup>+</sup> | .13<sup>c</sup> |
| TT for youth                        | $F(4, 2340) = 1.33$ | .257 | .45<sup>+++</sup> | (1.0) | .51<sup>+++</sup> | .38<sup>+</sup> | .11<sup>b</sup> |
| TT for parent                       | $F(4, 2340) = 2.21$ | .065 | .44<sup>b</sup><sup>+++</sup> | (0.0) | .33<sup>b</sup><sup>+++</sup> | (0.0) | --- |
| TT for teacher                      | $F(2, 2340) = 17.47$ | <.001 | --- | --- | --- | --- | --- |

| No treatment/ waitlist/ intervention | --- | --- | --- | --- | --- | --- |
| Psychotherapy placebo                | --- | --- | --- | --- | --- | --- |
| Pill placebo                         | --- | --- | --- | --- | --- | --- |
| Case management                      | --- | --- | --- | --- | --- | --- |
| Usual care treatment                 | --- | --- | --- | --- | --- | --- |

**Note.** Bonferroni correction was used ($p$ set at <.005) when testing two-way interactions for each clinical variable. Asterisks denote that the ES was significantly different from zero; identical letter superscripts within each row denote significant pairwise comparisons; a dash denotes zero studies in that subgroup; ES estimates in parentheses are based on fewer than five studies. ES = effect size; TP = target problem; TT = treatment type; CC = control condition.

<sup>f</sup>The behavioral categories included behavioral and cognitive-behavioral therapies. * $p < .05$. ** $p < .01$. *** $p < .001$. 

---

Lauren Krumholz Marchette
significant impact on ES. To assess whether these variables might relate to ES if combined, we aggregated them in two ways: (a) following Cochrane Collaboration guidelines (Higgins et al., 2011), we created a summary index for risk of bias, combining the six quality variables to classify each study as low risk, unclear risk, or high risk; and (b) we created a summary score for each study, ranging from 0 to 6, by summing across the six variables. Neither the summary index nor the summary score was significantly associated with ES.

Why the Drop in Mean ES?

The posttreatment mean ES of 0.45 in the current meta-analysis was substantially lower than the mean ESs (0.54 to 0.88) reported in all previous broad-based meta-analyses. We explored three possible explanations. One is that the many new studies added for the present meta-analysis showed more modest effects than prior studies. There is mixed evidence on this possibility: On one hand, study year was not significantly associated with ES (see Tables 1 and 4), but on the other hand, we found that studies in more recent decades were more likely to have larger samples and more rigorous control conditions, both characteristics associated with lower ES. A second possible explanation is that our use of multilevel modeling reduced ES in the present meta-analysis. This was not likely to be a major factor; multilevel models have been found to yield ES estimates comparable with traditional random-effects models (Van den Noortgate & Onghena, 2003). A third possible explanation is that differences in mean ES may result from changes over time in methods of synthesizing ES values. The older meta-analyses (from Casey & Berman, 1985; Kazdin et al., 1990; and Weisz et al., 1987) used an unweighted approach, but weighting ESs by the inverse of the sampling variance (see Hedges & Olkin, 1985) has subsequently become the norm in meta-analysis. Because larger ESs tend to have smaller weights (i.e., they tend to come from studies with

Brian C. Chu

Figure 3. Treatment Type × Informant Interaction showing that treatment effects depend, in part, on who reports the outcome. Effects were consistent and substantial across all three informants for youth-focused behavioral treatments (including CBT), but more variable across informants for other treatments. Two missing data points reflect the fact that teacher reports were not included in any study of caregiver/family nonbehavioral treatment or multisystem treatment. See the online article for the color version of this figure.
Table 4
Results of Moderator Analyses Based on Three-Level Mixed Effects Models of 1,802 Dependent ESs From 140 Studies at Follow-Up

<table>
<thead>
<tr>
<th>Study-level moderators (third level)</th>
<th>n of studies</th>
<th>n of ESs</th>
<th>Subgroup analysis</th>
<th>ES (g)</th>
<th>95% CI</th>
<th>Test statistic</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posttreatment lag time (weeks)</td>
<td>121</td>
<td>1,505</td>
<td></td>
<td></td>
<td></td>
<td>r(1503) = −1.32</td>
<td>.186</td>
</tr>
<tr>
<td>Study year</td>
<td>140</td>
<td>1,802</td>
<td></td>
<td></td>
<td></td>
<td>r(1800) = −1.08</td>
<td>.281</td>
</tr>
<tr>
<td>Study location</td>
<td>140</td>
<td>1,802</td>
<td></td>
<td></td>
<td></td>
<td>r(1800) = .55</td>
<td>.585</td>
</tr>
<tr>
<td>North America</td>
<td>100</td>
<td>1,431</td>
<td></td>
<td>.34***</td>
<td>.24</td>
<td>F(2, 1752) = .75</td>
<td>.475</td>
</tr>
<tr>
<td>Outside North America</td>
<td>40</td>
<td>371</td>
<td></td>
<td>.40***</td>
<td>.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant engagement</td>
<td>129</td>
<td>1,755</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recruited</td>
<td>81</td>
<td>1,237</td>
<td></td>
<td>.42***</td>
<td>.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referred</td>
<td>32</td>
<td>396</td>
<td></td>
<td>.29***</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonvoluntary</td>
<td>16</td>
<td>122</td>
<td></td>
<td>.37**</td>
<td>.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>82</td>
<td>1,086</td>
<td></td>
<td>.39***</td>
<td>.26</td>
<td>r(1084) = −.26</td>
<td>.7889</td>
</tr>
<tr>
<td>Caucasian sample (≥50% Caucasian)</td>
<td>65</td>
<td>915</td>
<td></td>
<td>.39***</td>
<td>.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Caucasian sample (&lt;50% Caucasian)</td>
<td>17</td>
<td>171</td>
<td></td>
<td>.35**</td>
<td>.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>127</td>
<td>1,670</td>
<td></td>
<td></td>
<td></td>
<td>r(1668) = .37</td>
<td>.708</td>
</tr>
<tr>
<td>Majority male (&gt;50% male)</td>
<td>81</td>
<td>1,234</td>
<td></td>
<td>.36***</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Majority female (&gt;50% female)</td>
<td>46</td>
<td>436</td>
<td></td>
<td>.39***</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developmental period</td>
<td>139</td>
<td>1,801</td>
<td></td>
<td></td>
<td></td>
<td>r(1799) = −1.42</td>
<td>.155</td>
</tr>
<tr>
<td>Childhood (mean age &lt;12 years)</td>
<td>87</td>
<td>1,356</td>
<td></td>
<td>.40***</td>
<td>.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adolescence (mean age ≥12 years)</td>
<td>52</td>
<td>445</td>
<td></td>
<td>.28***</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis requirement</td>
<td>68</td>
<td>986</td>
<td></td>
<td></td>
<td></td>
<td>r(984) = .12</td>
<td>.901</td>
</tr>
<tr>
<td>Required of all participants</td>
<td>45</td>
<td>713</td>
<td></td>
<td>.37***</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not required</td>
<td>23</td>
<td>273</td>
<td></td>
<td>.39***</td>
<td>.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted problem</td>
<td>140</td>
<td>1,802</td>
<td></td>
<td></td>
<td></td>
<td>F(4, 797) = 3.32</td>
<td>.010</td>
</tr>
<tr>
<td>ADHDa</td>
<td>28</td>
<td>492</td>
<td></td>
<td>.22*</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conduct</td>
<td>46</td>
<td>550</td>
<td></td>
<td>.44***</td>
<td>.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxietya,b</td>
<td>34</td>
<td>419</td>
<td></td>
<td>.55***</td>
<td>.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depressionb</td>
<td>27</td>
<td>295</td>
<td></td>
<td>.22*</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple problemsa,c</td>
<td>5</td>
<td>46</td>
<td></td>
<td>.02</td>
<td>−.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study-level moderators (second level)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informant</td>
<td>116</td>
<td>1,241</td>
<td></td>
<td></td>
<td></td>
<td>F(2, 1238) = 4.31</td>
<td>.014</td>
</tr>
<tr>
<td>Youtha</td>
<td>66</td>
<td>504</td>
<td></td>
<td>.33***</td>
<td>.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parentb</td>
<td>62</td>
<td>522</td>
<td></td>
<td>.36***</td>
<td>.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teachera,b</td>
<td>36</td>
<td>215</td>
<td></td>
<td>.21*</td>
<td>.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment typea,b</td>
<td>104</td>
<td>1,173</td>
<td></td>
<td></td>
<td></td>
<td>F(3, 1169) = .36</td>
<td>.782</td>
</tr>
<tr>
<td>Youth-focused behavioral</td>
<td>73</td>
<td>764</td>
<td></td>
<td>.37***</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth-focused nonbehavioral</td>
<td>8</td>
<td>46</td>
<td></td>
<td>.29*</td>
<td>.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caregiver/Family-focused behavioral</td>
<td>24</td>
<td>292</td>
<td></td>
<td>.42***</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multisystem</td>
<td>8</td>
<td>71</td>
<td></td>
<td>.40*</td>
<td>.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control condition†</td>
<td>138</td>
<td>1,794</td>
<td></td>
<td></td>
<td></td>
<td>F(3, 1790) = .18</td>
<td>.911</td>
</tr>
<tr>
<td>No treatment/waitlist</td>
<td>58</td>
<td>661</td>
<td></td>
<td>.39***</td>
<td>.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy placebo</td>
<td>43</td>
<td>715</td>
<td></td>
<td>.38***</td>
<td>.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case management</td>
<td>10</td>
<td>71</td>
<td></td>
<td>.36*</td>
<td>.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usual care treatment</td>
<td>35</td>
<td>347</td>
<td></td>
<td>.31***</td>
<td>.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Some moderators were missing for certain studies. Each study can contribute multiple ESs; thus, study sample size across subgroups can exceed the total study sample size for the ES-level moderators. Within each moderator having more than two subgroups, identical letter superscripts indicate significant (p < .05) pairwise comparisons between subgroups. ES = effect size; g = Hedges’ g; CI = confidence interval; ADHD = attention-deficit hyperactivity disorder.

† The behavioral categories included behavioral and cognitive-behavioral therapies. ‡ The pill placebo (n = 2) and caregiver/family-focused nonbehavioral (n = 2) categories were excluded because they contained fewer than five studies, and parameter estimates are poor when number of studies is very small. \* p < .05. ** p < .01. *** p < .001.

smaller sample sizes, yielding larger sampling variances), weighting can lower mean ES estimates. To illustrate, Weisz et al. (1995) found that switching from an unweighted to a weighted approach using the same data set reduced mean ES from 0.71 to 0.54. This difference suggests that failure to use weighting is a particularly credible factor in explaining the relatively large ES values found in the earliest meta-analyses. Notably, CI overlap indicates that the present ES mean of .45 (95% CI [.40, .50]) did not differ significantly from the mean ES of .54 (95% CI [.45, .63]) in Weisz et al., which did use ES weighting. By contrast, formal comparisons with the older ES values (from Casey & Berman, 1985; Kazdin et al., 1990; Weisz et al., 1987) are not meaningful, given the unweighted approach used in those meta-analyses. Thus, among possible reasons for the decline in mean ES, the shift to weighting ESs by the inverse of the sampling variance is especially plausible.
Discussion

By synthesizing 50 years of cumulative knowledge using a rigorous multilevel approach, this meta-analysis provided a particularly comprehensive and nuanced picture of the youth therapy evidence base, addressing some key questions for the field:

1. What is the overall effect of youth psychological therapy? We found a significant overall posttreatment ES of 0.46, which dropped nonsignificantly to 0.36 in follow-up assessments averaging 11 months later. Thus, five decades of research on youth psychological therapy points to beneficial effects that are moderate in magnitude and relatively durable. The posttreatment mean ES of 0.46 approached Cohen’s (1988) threshold of 0.5 for a medium effect, but was lower than the mean ESs (0.54 to 0.88) obtained in all previous broad-based meta-analyses (Casey & Berman, 1985; Kazdin et al., 1990; Weisz et al., 1987, 1995). We considered three explanations for the drop in mean ES relative to the earliest meta-analyses. Our analysis suggested that the most plausible explanation may be the emergence of the standard practice of weighting ESs by the inverse of the sampling variance (see Hedges & Olkin, 1985), a practice not employed in the earliest meta-analyses.

2. Does therapy impact differ by target problem? Yes, and markedly so. Target problem was the most potent moderator of treatment benefit, and the effect was not qualified by interactions with treatment type or control condition. At both posttreatment and follow-up, treatment effects were larger for anxiety than for other target problems, and consistently so across different informants. Candidate explanations might include the historically close tie between basic research and anxiety treatment methods (e.g., extinction research and exposure), the existence of a dominant path to intervention success (i.e., graduated exposure), the availability of an objectively observable measure of outcome (i.e., behavioral approach), and the high level of motivation for change, and the treatment compliance, so often seen in anxious youths (see Weisz, 2004). Across the four specific targeted problems, treatment of depression showed the most disappointing effects; in fact, by teacher report (see Tables 2 and 3), treatments for depression actually fared worse than control conditions. This finding is noteworthy in light of professional guidelines that recommend psychological therapy as the first-line treatment for youth depression (Birmaher et al., 2007; A. H. Cheung et al., 2007). We checked whether the relatively weak effects might reflect disproportionate use of more rigorous control conditions in depression studies, but in fact no-treatment/waitlist was by far the most common control condition in those studies, and the Target Problem × Control Condition interaction did not approach significance. Of course there might be other methodological explanations for the disappointing effect of depression treatments, but we could not identify any reason to discount the finding. Together with previous evidence of weak effects in youth depression treatment (Weisz, Jensen-Doss, & Hawley, 2006), the findings suggest that depression may be an appropriate priority for future treatment development and evaluation.

The target problem findings showed that treatment of multiple problems concurrently (i.e., within the same treatment episode) produced strikingly smaller mean ES than treatment of any single targeted problem, an effect that was not significantly different from zero at posttreatment or follow-up. It is possible that outcome assessment may be more complicated in multiproblem treatment, and this might have influenced the ES for this approach. It should also be noted that the finding was based on only 10 studies, so more evidence is needed. However, these studies combined do suggest that initial efforts to treat multiple problems concurrently have been less effective than focusing more narrowly (see Craske et al., 2007). To the extent that studies of a target problem exclude comorbidities,
this may be a concern because the evidence indicates that co-occurring problems are pervasive in clinically referred youths (Angold et al., 1999; Costello et al., 2003). The weak effects for multi-problem treatment may suggest a need for new ways to address youth comorbidity (see Barlow et al., 2010; Bearman & Weisz, 2015; Ehrenreich-May & Chu, 2013; Weisz, Krumholz, Santucci, Thomassin, & Ng, 2015).

3. Does therapy impact differ according to the therapy used? Our findings offer mixed evidence, with no overall treatment type moderator effect, but an intriguing interaction between treatment type and informant. The interaction, when broken down into component tests, showed that youth-focused behavioral treatments (including CBT) produced the most robust cross-informant evidence of beneficial effects (see Figure 3). These youth-focused behavioral therapies were alone among the intervention categories in showing significant effects across youth, parent, and teacher informants (see Weiss & Weisz, 1995; Weisz et al., 1995). That said, effects of other treatment approaches were significant according to some informant reports, producing a mixed picture. The absence of across-the-board superiority for behavioral treatments, combined with the moderate mean ES for caregiver/family-focused nonbehavioral treatments, certainly warrants attention in future research. Taken together, the findings suggest that different treatments may differ in their effects, but in highly specific ways that require precise analysis and close attention to the source of outcome information. Ideally, such analysis should include direct comparisons between active competing treatments, but such comparisons are few in number for most pairwise comparisons. This argues for triangulating strategies, complementing treatment–treatment comparisons with treatment–control meta-analyses that test treatment type and its interaction with other candidate moderators, as we have done here.

4. Does therapy impact depend on the control condition employed? Because RCTs and meta-analyses emphasize interventions and their effects, the control condition employed is often a kind of stealth moderator. Our findings suggest that this potent factor should be brought out of hiding. We found that treatments looked stronger when compared with relatively inert control conditions (e.g., wait-list) than when compared with active controls (see Kazdin et al., 1990), and comparisons with usual care generated the smallest treatment effects in both posttreatment and follow-up comparisons, with only one exception: Pill placebo control conditions produced smaller ES in follow-up tests, but this was based on an N of only two studies. Recent arguments have stressed the ecological validity of usual care control conditions and the importance of testing whether treatments can improve on current practice in clinical care (Spielmans et al., 2010; Weisz et al., 2013); the present findings suggest, further, that usual care may be a particularly rigorous standard of comparison.

5. Does it matter who reports the outcome? Evidently, it matters a lot. “Informant” was a remarkably pervasive moderator, showing significant effects in both posttreatment and follow-up analyses, and entering into multiple interactions with other moderators. This finding is consistent with evidence that interinformant agreement is modest, at best, on many measures of youth functioning and mental health (De Los Reyes et al., 2015). We found that youths and parents, the informants most likely to know the youth well and to know the treatment condition, generally reported larger effects than teachers; but the informant effect differed according to the target problem, treatment, and control condition. For example, there was no significant effect of informant on outcomes in anxiety treatment, which showed substantial ES for all three informants, but a strong (p < .001) informant effect on outcomes in depression treatment, which showed substantial positive effects according to...
Youth self-reports but negative effects (with depression treatments faring worse than control conditions) according to teachers. The fact that there were more effects involving informant than any other moderator underscores the need for both researchers and clinicians to obtain outcome information from multiple informants and to characterize outcome reports according to their source. The findings highlight the fact that youth therapy outcome is always, to some extent, in the eye of the beholder, and that different informants observe different samples of a youth’s behavior, in different contexts, and bring different perspectives to what they observe. This being the case, much could be gained by linking therapy research and practice to the burgeoning effort to build multi-informant approaches to youth mental health assessment (De Los Reyes et al., 2015).

Recent perspectives in both scientific (e.g., Drysdale et al., 2014) and popular media (e.g., Friedman, 2014) suggest the hypothesis that adolescents might show poor therapy response compared with other age groups. Some evidence suggests that adolescent neurological, social, and lifestyle factors may undermine emotion regulation and hamper attentional focus, and adolescent autonomy striving might undermine cooperation with a therapist and completion of therapy homework (Drysdale et al., 2014). However, our findings did not show majority adolescent samples to be significantly less responsive to therapy than majority child samples. This conclusion is consistent with a recent review by Kendall and Peterman (2015) focused on anxiety treatment, and with findings of separate meta-analyses focused on treatment of depression (Weisz et al., 2006) and disruptive behavior (Lundahl et al., 2006; and see mixed findings in S. Reynolds et al., 2012). The fact that well-developed therapies often use approaches designed to fit youth developmental level might mitigate age differences in outcome. In any event, our evidence does not indicate that psychological therapies have been less successful with adolescents than with children.

Two other findings warrant mention. First, study year was not associated with ES. This finding is consistent with other reports of little historical increase in treatment benefit (e.g., Johnsen & Friberg, 2015; Weisz et al., 2006). At first blush, this might be read as evidence that treatments have not become more effective, but multiple alternative interpretations merit consideration. It is possible that true increases in treatment efficacy over time may have been masked by temporal changes in the nature and design of research. For example, a more mature evidence base may be likely to include increased numbers of replication and extension studies rather than initial discovery trials. ESs for high impact discovery trials tend to be larger, especially if the initial studies were done with small samples (Ioannidis, 2005). Our data suggested that time trends in sample size might also have played a role. Across our study pool at posttreatment, mean group sample size grew larger over time (mean group n from the first to the last decade was 13.53, 12.01, 18.30, 35.80, and 37.67, respectively), F(4, 5134) = 11.80, p < .001; and larger study group samples were associated with smaller ES, r(5137) = −2.46, p = .014. Other aspects of study design may warrant attention as well. For example, increases over time in the use of more rigorous control conditions could affect time trends in ES. Use of the four types of control conditions did differ across the five decades, χ²(16) = 564.66, p < .001, with a striking increase in the use of usual care in the most recent two decades. Usual care may often include active treatment ingredients; in the present meta-analysis, it is the control condition associated with the lowest ES in most comparisons. That said, it is fair to note that our analyses revealed no significant effect of the interaction between study year and control condition on ES. Our data do not provide a definitive explanation for the absence of a significant increase in ES over the years, but multiple methodological explanations should be explored. If intensive examination were to reveal no viable artifactual or methodological explanation, then the flattening of ESs over time might suggest a need to rethink the very research strategy through which psychological therapies for youths have been developed across five decades.

Our ethnicity moderator tests found no significant difference in treatment benefit between majority Caucasian samples and majority non-Caucasian samples. This finding is consistent with findings of a recent review (S. J. Huey, Tilley, Jones, & Smith, 2014) indicating that psychological therapy is efficacious for ethnic minority youths and adults across multiple problem areas, and about equally efficacious for minorities and Caucasians. Although our findings are in harmony with those results, enthusiasm is tempered by the fact that many studies failed to report sample ethnicity, and that minorities were underrepresented; only 12.08% of our study pool reported a majority non-Caucasian sample. With more than 45% of American youths now ethnic minorities (Mather, Pollard, & Jacobson, 2011), youth therapy research is in need of more diverse samples, paired with full reporting of race/ethnicity and related results, if clinical science is to properly inform clinical practice.

Limitations of our findings reflect, in part, trade-offs associated with the inherent nature of meta-analysis (e.g., forming comparison groups by lumping fine-grained categories together) and the five-decade span of our study collection. Differences across the decades in study methods, reporting conventions, and author access limited prospects for fully addressing certain questions. For example, reporting limitations in the studies ruled out fine-grained analysis of race and ethnicity; and changes over the decades in
diagnostic categories, and in the nature and categories of professional clinical training, ruled out meaningful analyses of those factors (cf. Levenson, 2014). Some of the limitations might be addressed in the future by narrowing the focus and the range of study years, with the trade-off that the data set would then be less representative of the full body of youth treatment outcome research. Some limitations reflect the design of studies in the field, thus highlighting research needs for the future (e.g., the fact that so few studies complemented youth and parent-report outcome measures with teacher report underscores the need for more unbiased informants). Another limitation is that focusing our search on English language reports may have constrained the picture of youth therapy effects, particularly in light of our finding that posttreatment ES was larger in studies conducted outside the United States than in those within the United States. In addition, excluding unpublished dissertations may have limited the representativeness of our findings, potentially eliminating some studies with relatively low ES (see McLeod & Weisz, 2004). These limitations might be usefully addressed in future meta-analyses. In addition, future work could be enriched by comparing an array of treatment outcome assessment methods that are increasingly a part of randomized trials, including behavioral observations, performance tasks, and independent evaluator assessments of symptoms and functioning. Finally, the fact that our findings left substantial amounts of variance in treatment efficacy unexplained strongly suggests that influential moderators not identified here remain to be discovered in future research.

As noted, findings from the past 50 years suggest both benefits of psychological therapy and important directions for the future in youth therapy research and clinical practice. Perhaps the broadest implication derives from translating our overall mean ES of 0.46 at posttreatment into a “common language effect size” (McGraw & Wong, 1992): The probability that a randomly selected youth in the treatment condition would be better off after treatment than a randomly selected youth in the control condition was 63%—only moderately better than chance at 50%. This suggests that, although youth therapies have produced beneficial effects over the years, there is much room for improvement—ample opportunity for clinical scientists and practitioners, working together, to strengthen clinical care for young people and their families, who deserve the best interventions our collaborative efforts can produce.

References

References marked with an asterisk indicate studies included in the meta-analysis.


Ahrens, J., & Rexford, L. (2002). Cognitive processing therapy for incarcerated adolescents with PTSD. Journal of Aggression, Maltreatment & Trauma, 6, 201–216. http://dx.doi.org/10.1300/J146v06n01_10


104 WEISZ ET AL.


This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.


The document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.


Weisz et al.  

110


This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.


*Treatment for Adolescents With Depression Study Team. (2003). Treatment for Adolescents With Depression Study (TADS): Rationale, de-


Received March 21, 2015
Revision received December 14, 2015
Accepted December 16, 2015