Using Dissertations to Examine Potential Bias in Child and Adolescent Clinical Trials

Bryce D. McLeod and John R. Weisz
University of California, Los Angeles

The major youth psychotherapy meta-analyses have relied on published studies, which may have led to biased effect size estimates. To examine this possibility, the authors compared 121 dissertations with 134 published studies and found the following: (a) few differences on individual methodological variables, but, overall, stronger methodology in dissertations; (b) no differences in the steps taken to ensure treatment integrity; and (c) a mean dissertation effect size less than half that of published studies. The effect size difference remained robust across tests controlling for all reliable method differences. The findings suggest that dissertations are so strong, both methodologically and clinically, that they warrant inclusion in child psychotherapy meta-analyses and that previous meta-analyses, by excluding them, may have overestimated treatment effects.

Over the past 18 years, four broad-based meta-analyses have expanded our knowledge about child and adolescent psychotherapy effects (Casey & Berman, 1985; Kazdin, Bass, Ayers, & Rodgers, 1990; Weisz, Weiss, Aliche, & Klotz, 1987; Weisz, Weiss, Han, Granger, & Morton, 1995). These four meta-analyses have demonstrated that across a variety of treatments for a number of different problems, therapy for children and adolescents (herein referred to collectively as “children”) produces mean effect size (ES) values in the medium to large range (range: .71 to .88; see Cohen, 1988): The consistency of these findings, based on over 350 independent studies, appears to support the efficacy of child psychotherapy.

An important caveat is that these findings are only as representative of treatment outcome research as their collections of studies are. The representativeness of these meta-analyses may be questioned because all four included only published studies. Relying on published work does provide quality control, ruling out studies that could not survive scrutiny by reviewers and editors. However, relying exclusively on published work can also introduce distortion because of the combined effects of publication bias and the file drawer effect. Ample evidence indicates that journal editors and reviewers favor publication of statistically significant findings (i.e., publication bias—see, e.g., Atkinson, Furlong, & Wampold, 1982; Coursol & Wagner, 1986) and that authors tend to file away research that does not generate statistically significant results (i.e., file drawer effect—see, e.g., Rosenthal, 1979; Sterling, 1959). Through the joint effects of publication bias and the file drawer effect, meta-analyses that rely solely on published studies may generate overestimates of mean ES (Begg, 1994; Lipsey & Wilson, 1993; Smart, 1964).

Paradoxically, the fact that the four meta-analyses relied on published studies to ensure methodological quality may actually have increased the likelihood that the meta-analytic findings would misrepresent child psychotherapy outcome findings. To adequately investigate this possibility, however, a representative sample of methodologically sound and clinically strong studies is needed whose findings can be compared with the published literature. Dissertations may constitute such a sample. Including dissertations in meta-analyses may help meta-analysts obtain a representative sample of studies and, in turn, provide a more accurate ES estimate. Of course, dissertations will only help generate a more accurate ES estimate if they are comparable to published studies in strength of research methods and treatment procedures. One could question whether dissertations are as strong in research methodology as published studies, but the dissertation process, guided by a faculty advisor and committee and marked at beginning and end by an oral exam, may help ensure quality control. The methodological quality of dissertations can certainly be addressed as an empirical question. Another concern may be that dissertations test treatment procedures that are less substantial, but this concern can also be made an empirical question. So, dissertations may provide a valuable resource that can help provide a more accurate ES estimate, if it can be established that they are not weaker, methodologically or clinically, than published studies.

1 Rosenthal (1979) may have been the first to propose a method to account for the file drawer effect. He developed a statistical formula to calculate how many null findings would have to exist in file drawers to invalidate the effect estimate of a meta-analysis. Although this formula can gauge the amount of threat the file drawer phenomenon may pose to a meta-analysis, the formula cannot provide an indication of the effects within the population.
In the present study, we aggregated outcome studies identified through the Dissertation Abstracts International (DAI) database and assessed their comparability in research design and methodology and in the strength and quality of the clinical interventions tested to a sample of published studies. Next, we compared mean ES across the two pools of studies, with methodological and clinical variables controlled, to examine whether a focus on dissertations might alter accepted conclusions about the effects of child and adolescent therapy.

**Method**

We used the same definition of “psychotherapy” used in the prior meta-analyses by Weisz et al. (1987, 1995); any intervention designed to alleviate psychological distress, reduce maladaptive behavior, or enhance adaptive behavior through counseling, structured or unstructured interaction, a training program, or a predetermined treatment plan. As in Weisz et al. (1995), treatments were excluded if they only included reading interventions (e.g., bibliotherapy), involved drugs in any step of the treatment, used teaching or instruction intended only to increase knowledge of specific subjects, or targeted children considered to be at risk for the sole purpose of preventing the development of problems.

**Literature Search**

Dissertations were used to represent completed studies not subject to publication bias or the file drawer effect. We used three different methods to identify dissertations. First, we used the DAI computer database, which indexes (with key terms) and abstracts all dissertations completed from 1980 to date. The search covered January 1980–December 1999, and we used the same 21 psychotherapy-related key terms and synonyms used by Weisz et al. (1987, 1995): client-centered, contract-(- indicates ing, systems, etc.) counseling, cotherapy, dream analysis, insight-, intervention-, model-, modifies, operators, paradox, psychoanalysis, psychodrama-, psychotherapy-, reinforce-, respondent, role-playing, therapy-, training, transactional, and treatment. Also as in Weisz et al. (1987, 1995), these terms were crossed with the following age group and topic constraints: adolescent-, child-, juvenile-, pre-adolescent-, and youth-, and assess-, comparison, effect-, efficacy, evaluat-, influence, impact, and outcome. Second, we searched relevant child psychotherapy meta-analyses (e.g., Baer & Nietzel, 1991; Beelmann, Pfingsten, & Loesel, 1994; Ceder & Levant, 1990; Durlak, Fuhrman, & Lampman, 1991; Dush, Hirt, & Schroeder, 1989; Graziano & Diamant, 1992; Grossman & Hughes, 1992; Haney & Durlak, 1998; Hoag & Burlingame, 1997; Reinecke, Ryan, & DuBois, 1998; Skiba & Casey, 1985) to identify relevant dissertations. Third, we searched the microfilm versions of DAI dated 1980 to 1999 to locate relevant dissertations. These steps produced an initial pool of 1,824 dissertations, which were reduced in a stepwise fashion using title, abstract, method section, and results section to produce a pool of dissertations that met inclusion requirements.

**Design and Reporting Requirements**

To be included, a dissertation had to meet certain basic methodological requirements. First, participants had to be randomly assigned to groups. Nonrandom assignment, of course, makes findings difficult to interpret with confidence. In addition, studies that use random assignment have been found to generate higher, less variable ES than studies that do not (Greenwald & Russell, 1991; Lipsey & Wilson, 1993; Shadish & Ragdell, 1996; Weiss & Weisz, 1990); this inclusion criterion thus helped ensure a conservative test of whether dissertations showed smaller effects than published studies. Second, as in all four previous broad-based meta-analyses (Casey & Berman, 1985; Kazdin et al., 1990; Weisz et al., 1987, 1995), we excluded single-subject and within-subject designs. ES for such studies is based on intraparticipant variance; design and computational differences make it difficult to fairly combine these studies with studies that include independent treatment and control groups. Finally, as in all four previous meta-analyses, we excluded studies that reported only follow-up data, with no immediate posttreatment data.

**Dissertation Outcome Studies Generated by the Search**

The reduction process led to a pool of 121 dissertation psychotherapy outcome studies (in 113 dissertations) with 199 different treatment groups. The dissertations were completed between 1980 and 1999. The mean age of treated youngsters ranged from 2.32 to 18.50 years ($M = 11.2, SD = 3.37$).

**Comparison Pool of Published Studies From the Weisz et Al. (1995) Meta-Analysis**

For statistical tests comparing dissertations and published studies from the meta-analytic literature, we used studies from the most recent broad-based child meta-analysis, by Weisz et al. (1995). With only random assignment studies included, the total was 107 published studies (reported in 106 articles). However, the year of publication for the published study set ranged from 1964 to 1993, whereas the year of publication for the dissertation study set ranged from 1980 to 1999. It was therefore necessary to remove studies from the published study set published prior to 1980 (i.e., 7 studies) and update the published study set to 1999 to ensure that the two study sets encompassed comparable time periods. We used two different approaches to identify relevant published studies to update the study set. First, we conducted a computer search for the period January 1994 through December 1999, using the same 21 psychotherapy-related key terms mentioned previously. Second, still focusing on the same time period, we also searched by hand, issue by issue, journals that had produced articles in our previous meta-analyses. These steps produced a pool of 34 studies, bringing the total to 139 published studies (reported in 135 articles). We next used PsycLIT and contacted authors to check for overlap between dissertations and our pool of published studies. Studies based on 5 of the dissertations (Coats, 1982; Denkowski, 1981; Kaduson, 1993; Larson, 1984; Stark, 1985) were included in our collection of 139 published studies. To ensure that dissertations and published studies in our two sets represented independent groups, we dropped these 5 from our pool of published studies. This left 134 published studies (reported in 129 articles), which included 211 different treatment groups. Year of publication ranged from 1980 to 1999. Mean age of the samples ranged from 1.5 to 17.6 years ($M = 10.58, SD = 3.59$).

**Coding of the Studies**

The same codes for therapy methods, treated problems, and study design features described in Weisz et al. (1995) were used for dissertations and published studies. Additional codes focusing on methodological and treatment characteristics were devised. All codes were classified into five domains according to existing typologies (see Cook & Campbell, 1979; Kazdin, 1994; Shapiro & Shapiro, 1983): internal validity, external validity,

---

2 Comparisons between the two study sets were run with the overlapping studies removed from both study sets, and published studies still produced significantly higher ES estimates than dissertation studies (all $ps < .006$).

3 Because the idea of “psychotherapy” with children younger than 2 years may be puzzling, we should note that studies focused on problems of young children tended to involve parent training.
ity, construct validity, methodological characteristics, and treatment characteristics.

**Internal validity dimensions.** Each study was coded on three internal validity dimensions: (a) experimental attrition, (b) participant assignment (randomly vs. randomly with matching of groups on at least one dependent variable), and (c) type of control group (assessment only, waiting list, attention-placebo, or minimal treatment).

**External validity dimensions.** Published studies and dissertations were coded on two external validity dimensions: (a) therapist type (professionals, graduate and/or professional students, or paraprofessionals) and (b) treatment setting (clinical vs. nonclinical setting).

**Construct validity dimensions.** Every study was coded on four construct validity dimensions: (a) measurement technology of outcome measure (self-report data, subjective scales, objective counts, and life-event data), (b) source of outcome measure (observers, therapists, parents, participants’ self-reports, participants’ own performance, expert judges, peers, teachers, and interviewer ratings), (c) rater blindness, and (d) participant blindness.

**Methodological characteristics.** Every study was also coded on two methodological characteristics: (a) sample size and (b) number of outcome measures.

**Treatment characteristics.** Published studies and dissertations were coded on the following 13 treatment characteristics: (a) treatment format (individual vs. group, number of therapists, and number of clients), (b) treatment contact (amount of therapist contact with the child, parents family, and/or teacher), (c) treatment dose (number of sessions, length of sessions, total time spent in therapy), (d) theoretical orientation (behavioral vs. nonbehavioral), (e) target problem (controlled vs. overcontrolled), and (f) treatment integrity (pretherapy training, treatment manual, therapist supervision, and adherence checks).

**Interrater Agreement**

Assessment of interrater agreement progressed through two stages. In Stage 1, we used the same codes used by Weisz et al. (1995) for dissertations and published studies. These codes had been used reliably in Weisz et al., but we also had two raters, following training in the coding system, independently code a sample of 28 studies. In Stage 2, additional codes were devised. Because these were new codes, we assessed reliability for a substantial sample, with two coders, following training in the coding system, independently coding a sample of 50 studies. Following Fleiss (1981), kappas below .40 reflect “poor” agreement, kappas between .40 and .74 reflect “fair to good” agreement, and kappas of .75 and higher reflect “excellent” agreement. Intercoder agreement ranged from .64 to 1.00, with 6 of our codes in the fair to good range and 17 in the excellent range.

**Results**

We began by testing whether dissertations and published studies differed in the quality of their research methods or the characteristics and quality of their treatment procedures. We then tested whether these two study sets differed in mean ES. Finally, we assessed whether various methodological and statistical factors might account for the ES difference between the two study sets.

**Do Dissertations and Published Studies Differ Methodologically or Clinically?**

To compare the 121 dissertation studies to the 134 published studies on methodological factors and treatment characteristics, we included tests with all 25 of the variables listed under the Coding of the Studies heading. We also created three composite variables to reflect methodological quality, amount of treatment contact, and treatment integrity.

**Methodological quality.** Two approaches have been used in the meta-analytic literature to quantify the methodological quality of individual studies: (a) construction of a composite score in which the method of participant assignment is heavily weighted (random vs. no random assignment to groups; see Lipsey & Wilson, 1993) and (b) division of studies into high-, medium-, and low-methodological-quality categories (see Lipsey & Wilson, 1993). There is, however, no consensus as to which approach is preferable or even as to which methodological factors should be included (Chambless & Hollen, 1998; Chambless & Ollendick, 2000; Jüni, Witschi, Bloch, & Egger, 1999; Kazdin, 1994, 1999).

Lacking clear guidelines regarding the methodological characteristics of “high-quality” treatment outcome studies, we opted to construct a composite index of methodological quality, as opposed to a categorical approach (i.e., designating studies low, medium, and high quality), comprising design and methodological factors considered important within the treatment outcome literature (Chambless & Hollen, 1998; Chambless & Ollendick, 2000; Durlak, Wells, Cotton, & Johnson, 1995; Kazdin, 1994, 1999). The composite, methodological quality, was created by awarding 1 point for superiority on each of nine methodological factors: participant assignment (1 point for matching of treatment and control groups on participant characteristics) prior to randomization to groups, see Kazdin, 1994, 1999, total attrition (1 point for less than 10% total attrition; see Chambless & Hollen, 1998; Chambless & Ollendick, 2000; Durlak et al., 1995; Kazdin, 1994, 1999), differential attrition (1 point for less than 10% differential attrition; see Chambless & Hollen, 1998; Hansen, Collins, Malotte, Johnson, & Fielding, 1985), sample size (1 point for studies that averaged at least 30 per group; see Chambless & Hollen, 1998; Durlak et al., 1991), type of control group (1 point for attention placebo or minimal treatment to control for experimenter attention; see Chambless & Hollen, 1998; Durlak et al., 1995; Kazdin, 1994, 1999), measurement technology (1 point for either data based on objective counts [e.g., observed occurrences of a target behavior] or real-life-event data [e.g., grade point average, arrests; see Weiss & Weisz, 1990]), source of outcome measure (1 point for outcome data from trained observers or from participants’ actual performance on target tasks; see Weiss & Weisz, 1990), rater blindness (1 point if raters did not know of the participant’s assignment to treatment or control group; see Chambless & Hollen, 1998; Chambless & Ollendick, 2000), and participant blindness (1 point if participants were not aware that outcome assessment was being conducted; see Chambless & Hollen, 1998; Chambless & Ollendick, 2000).
Of these nine factors, only sample size, differential attrition, and participant assignment showed a significant difference. Published studies used somewhat larger samples than dissertations (M = 19.23, SD = 16.19 vs. M = 13.20, SD = 9.02 per group), $F(1, 406) = 21.17, p < .01$. Dissertations had a significantly lower rate of differential attrition (M = 0.03, SD = 0.06) compared with published studies (M = 0.05, SD = 0.08), $F(1, 393) = 7.84, p < .01$. Participants were randomly assigned with matching to groups in 9.7% of the published studies versus 19.8% in dissertations, $\chi^2(1, N = 254) = 5.26, p < .02$. On the full composite measure of methodological quality, we found a significant difference, $F(1, 239) = 5.26, p < .02$, favoring dissertations (M = 3.55, SD = 1.27) over published studies (M = 3.21, SD = 1.34).

Table 1 shows these comparisons as well as others involving additional variables that were not indices of methodological quality but which describe the methods used in the studies. These other variables yielded one significant effect. Dissertations included more outcome measures than published studies (M = 11.02, SD = 12.22 vs. M = 5.81, SD = 4.45), $F(1, 254) = 23.00, p < .01$.

### Treatment contact
The second composite, treatment contact, was created in two steps. First, using the participant contact code, we awarded 0 to 1 point (in quarter point increments) for therapist contact with each of the following: child, parents, family, and/or teacher. Second, using the full distribution of all 255 studies on the total time in treatment variable (Number of Sessions × Length of...
Sessions), we awarded .25 points to the lowest quartile, .50 points for the second, .75 points for the third, and 1.00 point for the upper quartile. Scores on this composite could thus range from 0 to 5 points. Comparisons of dissertations and published studies on the length of session (in minutes) code showed a significant difference, $F(1, 235) = 4.78, p < .04$, with published studies using longer sessions ($M = 66.01, SD = 44.60$) than dissertations ($M = 53.50, SD = 48.02$). Scores on the full treatment contact composite also showed a significant difference, $F(1, 356) = 8.36, p < .01$, favoring published studies ($M = 1.78, SD = 0.51$) over dissertations ($M = 1.65, SD = 0.36$).

**Treatment integrity.** The third composite, treatment integrity, was created by awarding 1 point for superior quality on each of four treatment integrity factors: pretherapy training (1 point for providing pretherapy training to therapists), treatment manual (1 point for the use of a manual or detailed description of treatment procedures), therapist supervision (1 point for providing supervision), and adherence checks (1 point for using such checks). Separate tests of the four factors showed no significant group differences. In addition, there was virtually no difference between published studies ($M = 1.33, SD = 1.36$) and dissertations ($M = 1.34, SD = 1.19$) on the full treatment integrity composite.

Table 2 shows the findings on the treatment contact and treatment integrity measures and on additional treatment characteristics unrelated to amount or integrity of treatment. The additional comparisons showed three significant effects. First, professional ther-

### Table 2

Comparisons Between Published Literature and Dissertations on Treatment Characteristics

<table>
<thead>
<tr>
<th>Treatment variable</th>
<th>Published lit.</th>
<th>Dissertation</th>
<th>$p &lt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment dose–intensity factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant contact$^a$</td>
<td>1.09 (0.34)</td>
<td>1.05 (0.22)</td>
<td>ns</td>
</tr>
<tr>
<td>Total time in treatment (in min)$^b$</td>
<td>804.29 (865.01)</td>
<td>635.58 (1,051.9)</td>
<td>ns</td>
</tr>
<tr>
<td>No. of sessions$^a$</td>
<td>11.6 (8.15)</td>
<td>12.29 (18.03)</td>
<td>ns</td>
</tr>
<tr>
<td>Length of sessions (in min)$^a$</td>
<td>66.01 (44.6)</td>
<td>53.5 (48.02)</td>
<td>.04</td>
</tr>
<tr>
<td>Treatment contact composite$^a$</td>
<td>1.78 (0.51)</td>
<td>1.65 (0.36)</td>
<td>.001</td>
</tr>
<tr>
<td><strong>Treatment integrity factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretherapy training, %$^b$</td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>No</td>
<td>66.4</td>
<td>64.5</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33.6</td>
<td>35.5</td>
<td></td>
</tr>
<tr>
<td>Treatment manual, %$^b$</td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>No</td>
<td>46.4</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>53.6</td>
<td>57.6</td>
<td></td>
</tr>
<tr>
<td>Therapist supervision, %$^b$</td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>No</td>
<td>79.4</td>
<td>82.0</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>20.6</td>
<td>18.0</td>
<td></td>
</tr>
<tr>
<td>Adherence checks, %$^b$</td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>No</td>
<td>73.5</td>
<td>77.9</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>26.5</td>
<td>22.1</td>
<td></td>
</tr>
<tr>
<td>Treatment integrity composite$^a$</td>
<td>1.33 (1.36)</td>
<td>1.34 (1.19)</td>
<td>ns</td>
</tr>
<tr>
<td><strong>Other treatment characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual vs. group treatment, %$^b$</td>
<td></td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>Individual</td>
<td>33.3</td>
<td>21.3</td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>63.6</td>
<td>77.0</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>3.1</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>No. of clients$^a$</td>
<td>6.05 (4.52)</td>
<td>6.3 (4.39)</td>
<td>ns</td>
</tr>
<tr>
<td>No. of therapists$^a$</td>
<td>1.22 (0.46)</td>
<td>1.21 (0.41)</td>
<td>ns</td>
</tr>
<tr>
<td>Therapist type, %$^b$</td>
<td></td>
<td></td>
<td>.0001</td>
</tr>
<tr>
<td>Professional</td>
<td>34.7</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>Graduate/professional student</td>
<td>51.0</td>
<td>64.8</td>
<td>.04</td>
</tr>
<tr>
<td>Paraprofessional</td>
<td>14.3</td>
<td>25.4</td>
<td>.04</td>
</tr>
<tr>
<td>Theoretical orientation, %$^b$</td>
<td></td>
<td></td>
<td>.03</td>
</tr>
<tr>
<td>Behavioral</td>
<td>80.6</td>
<td>86.4</td>
<td></td>
</tr>
<tr>
<td>Nonbehavioral</td>
<td>13.7</td>
<td>12.6</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>5.7</td>
<td>1.0</td>
<td>.009</td>
</tr>
<tr>
<td>Target problem, %$^b$</td>
<td></td>
<td></td>
<td>.0001</td>
</tr>
<tr>
<td>Undercontrolled</td>
<td>35.5</td>
<td>59.7</td>
<td></td>
</tr>
<tr>
<td>Overcontrolled</td>
<td>33.3</td>
<td>27.4</td>
<td>.0001</td>
</tr>
<tr>
<td>Other</td>
<td>31.2</td>
<td>12.9</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Note. Significant differences reflect chi-square tests comparing published studies and dissertations or $F$ tests comparing means between these two groups. 
$^a$ The mean (and standard deviation) was calculated at the treatment group level. $^b$ The percentage reflects the proportion of treatment groups within each study set using each subcategory of the treatment characteristic.
apists were used more in published study treatment groups (34.7% vs. 9.8%), $\chi^2(1, N = 218) = 20.31, p < .01$, whereas dissertation treatment groups used more graduate and professional students (64.8% vs. 51%), $\chi^2(1, N = 218) = 4.23, p < .04$, and paraprofessionals (25.4% vs. 14.3%), $\chi^2(1, N = 218) = 4.13, p < .04$. Second, the $2 \times 3$ (Study Type $\times$ Treatment Type) table was significant, $\chi^2(2, N = 408) = 7.11, p < .03$. Tests for each theoretical orientation showed no differences for behavioral or nonbehavioral treatments, but “other” treatments (i.e., interventions that combined “behavioral” and “nonbehavioral” techniques) were used more often in the treatment groups of published studies than dissertations (5.7% vs. 1%), $\chi^2(1, N = 408) = 6.81, p < .01$. Finally, the $2 \times 3$ Study Type $\times$ Target Problem table was significant, $\chi^2(2, N = 260) = 18.54, p < .01$, and tests for each individual problem type showed that dissertations were more likely than published studies to target undercontrolled problems (59.7% vs. 35.5%), $\chi^2(1, N = 260) = 15.32, p < .01$, whereas dissertations were less likely than published studies to focus on problems in the “other” category (e.g., social relationship problems; 12.9% vs. 31.2%), $\chi^2(1, N = 260) = 12.48, p < .01$.

**Do Dissertations and Published Studies Differ in ES?**

Next we tested for ES differences between dissertations and published studies. To facilitate comparisons with past treatment outcome meta-analyses of published studies, we followed the same procedures used by Weisz et al. (1987, 1995) to calculate ES estimates. We computed an ES for each study by dividing each study’s posttherapy treatment group versus control group mean difference by the standard deviation of the control group.7 When information needed for this calculation was not reported, we tried to contact authors to obtain the missing data. If authors did not provide the data, following Weisz et al. (1995), we derived ES values from inferential statistics reported in the study, using procedures recommended by M. L. Smith, Glass, and Miller (1980). Throughout the calculations, outcome measures and treatment groups were collapsed up to the level of analysis. For example, treatment groups were collapsed except when the effect of therapy type was being tested.

To permit a thorough comparison of the published studies and dissertations, we analyzed the data with two different approaches used in previous meta-analyses: unweighted least squares (ULS) general linear model (GLM) analyses and a weighted least squares (WLS) GLM approach. ULS was used in all four past meta-analyses (Casey & Berman, 1985; Kazdin et al., 1990; Weisz et al., 1987, 1995), so using it here facilitated comparison with previous findings. Compared with the ULS approach, the WLS approach is now more widely used, and it offers advantages. In the WLS approach, each ES is weighted by the inverse of its variance (Hedges, 1994; Hedges & Olkin, 1985); the analyses adjust for heterogeneity of variance across individual observations. In the ULS analyses, following Weisz et al. (1995), who used WLS in addition to ULS, we dropped academic outcomes (because so many factors other than psychopathology and therapy can influence school performance) and outliers, thus permitting comparison with the previous findings. As in Weisz et al. (and following Bollen, 1989), outliers were operationally defined as ESs lying beyond the first gap of at least one standard deviation between adjacent ESs in a positive or negative direction. That is, with study ES values arrayed in order of size, whenever two adjacent studies showed an ES difference of equal to or greater than one standard deviation, the more extreme ES value and all that were more extreme were identified as outliers.

The overall mean ES, calculated with ULS, based on one mean ES for each of the 121 dissertation studies, was 0.27 (significantly different from 0), t(120) = 4.7, p < .01. ES values ranged from −0.71 to 4.40 (Mdn = 0.10, SD = 0.64; 95% confidence interval [CI]: 0.16, .39). As Figure 1 shows, the mean ES differs markedly from the ULS mean effects reported in the four previous child psychotherapy meta-analyses: 0.71 (SD = 0.73) in Casey and Berman (1985), 0.79 (SD = 0.88) in Weisz et al. (1987), 0.88 (SD = 0.74) and 0.77 (SD = 0.71) in Kazdin et al. (1990); for treatment vs. no-treatment and for treatment vs. minimal treatment control group comparisons, respectively), and 0.71 (SD = 0.96) in Weisz et al. (1995). Using the WLS method, again with only one mean ES per study, we found the mean ES across the 121 dissertation studies was 0.23 (SD = 1.26; 95% CI: 0.14, 0.32). The only one of the four previous broad-based child meta-analyses to have used WLS was Weisz et al.; the mean ES reported of which was 0.54 (SD = 1.51).

We compared ES for our dissertation sample (N = 121) and our sample of published studies (N = 134). The ULS analysis showed a significant difference in ES between the published studies (M = 0.63, Mdn = 0.49, SD = 0.75) and dissertations (M = 0.27, Mdn = 0.10, SD = 0.64), F(1, 254) = 16.08, p < .01. The WLS analysis also showed a significant difference between published studies (M = 0.50, SD = 1.64) and dissertations (M = 0.23, SD = 1.26), F(1, 254) = 16.35, p < .01.

**Methodological and Statistical Factors That Might Explain the ES Difference Between Published Studies and Dissertations**

**Methodological factors.** Next, we tested whether the ES difference between published studies and dissertations held when the effects of the variables on which the two study sets differed significantly (shown in Tables 1 and 2) were controlled. We used simultaneous WLS regression to determine whether publication status remained significant after the methodological differences between the two study sets were controlled. This analysis yielded a significant multiple $R^2 = .15, F(12, 303) = 4.36, p < .01$. Publication status remained significant ($\beta = .28, p < .01$), with published studies producing significantly higher ES estimates, and four other variables also remained significant. ES was significantly higher when (a) participants were matched prior to being randomly assigned to groups ($\beta = .14, p < .02$), (b) fewer outcome measures were used ($\beta = -.18, p < .01$), (c) therapy sessions were shorter.

---

7 In calculating ES values, some meta-analysts (e.g., Casey & Berman, 1985; Hedges, 1982) favor dividing by the pooled standard deviation of treatment and control group; however, other meta-analysts (e.g., Smith et al., 1980; Weisz et al., 1987, 1995) believe that dividing by the control group standard deviation is more appropriate, because one effect of treatment can be to make variability greater in the treatment group than in the control group. The latter position has been supported in previous analyses (see Weiss & Weisz, 1990; Weisz et al., 1995).
Methodologically weak studies. 

To help ensure that the presence of methodologically weak studies did not account for the observed ES difference, we performed a median split on the methodological quality composite and removed all studies below the median value \((Mdn = 3.38)\). We then compared the remaining published studies \((n = 48)\) and dissertations \((n = 73)\). The WLS analysis revealed that, with methodologically weak studies removed, the published studies \((M = 0.47, SD = 1.60)\) continued to show significantly higher ES than dissertations \((M = 0.27, SD = 1.33)\), \(F(1, 120) = 4.74, p < .03\).

Discussion

The study compared dissertations with published studies on 25 methodological variables and on composite measures of design outlined by Huffcutt and Arthur (1995), we identified outlier data points in each study set. This led to the removal of seven studies from the dissertation study set, reducing the mean ES from 0.27 \((SD = 0.64, Mdn = 0.10, Skewness = 3.21, Skewness SE = 0.22)\) to 0.16 \((SD = 0.36, Mdn = 0.06, Skewness = 0.82, Skewness SE = 0.23)\). The procedure removed 13 studies from the published study set, reducing the mean ES from 0.63 \((SD = 0.75, Mdn = 0.49, Skewness = 2.89, Skewness SE = 0.21)\) to 0.49 \((SD = 0.40, Mdn = 0.48, Skewness = 0.48, Skewness SE = 0.22)\). With the 20 outlier studies removed, a WLS analysis showed a significant difference between published studies \((M = 0.46, SD = 1.03)\) and dissertations \((M = 0.16, SD = 0.86)\), \(F(1, 234) = 43.39, p < .01\), indicating that the presence of extreme outliers did not account for the observed ES difference between published studies and dissertations.

Figure 1. Mean effect sizes for child and adolescent psychotherapy found in 10 meta-analytic calculations: (a) Casey and Berman (1985; unweighted least squares [ULS] method only: \(M = 0.71, SD = 0.73\)); (b) Weisz et al. (1987; ULS: \(M = 0.79, SD = 0.88\))); (c) Kazdin et al. (1990; ULS: \(M = 0.88, SD = 0.74\)) for treatment versus no-treatment comparisons; (d) Kazdin et al. (1990; ULS: \(M = 0.77, SD = 0.71\)) for treatment versus active control comparisons; (e) Weisz et al. (1995; ULS: \(M = 0.71, SD = 0.96\)); (f) Weisz et al. (1995; weighted least squares [WLS]: \(M = 0.54, SD = 1.51\)); (g) published studies in the present study with ULS \((M = 0.63, SD = 0.49)\); (h) published studies in the present study with WLS \((M = 0.50, SD = 1.64)\); (i) dissertations in the present study with ULS \((M = 0.27, SD = 0.64)\); (j) dissertations in the present study with WLS \((M = 0.23, SD = 1.26)\).
quality, amount of treatment contact, and treatment integrity. On most measures, the study sets were quite similar, and dissertations were superior to published studies on the composite methodological quality variable. When we assessed outcome findings, the dissertation studies showed effects less than half the magnitude of those reported in meta-analyses of the published literature and significantly and robustly smaller than those of a comparison sample of 134 published studies. The difference in effect magnitude held up across multiple tests controlling for method differences and potential artifactual explanations. Evidently, the effects generated in an unbiased sample of dissertation outcome studies are robustly lower than those reported in the published literature, and the difference cannot be explained by methodological or substantive differences between the two pools of studies (at least, not by the 29 variables investigated here). This suggests that published studies overestimate the effects of child psychotherapy and do so to a substantial degree.

In reaching this conclusion, we are not by any means suggesting that child and adolescent psychotherapy effects are completely illusory or completely explained by biasing effects or that treatment is without real benefit. On the contrary, even the relatively conservative estimates provided by our dissertation sample point to positive effects (ULS = 0.27; WLS = 0.23) in the small to medium range (following Cohen, 1988). The dissertation ES ranged from −0.49 to 4.40, and 63% of all the dissertation ES values were positive. Thus, our data indicate that, even in a sample of studies that is arguably free of publication bias and the file drawer phenomenon, the treatments tested generally showed beneficial effects. So, the “half-full” perspective on our findings is that they support the efficacy of treatment for children and adolescents by showing that effects are positive even when publication bias and the file drawer phenomenon are stripped away. The evidence of beneficial psychotherapy effects for young people, reported in previous meta-analyses, thus cannot be dismissed as entirely an artifact of bias.

Given what appears to be a very substantial susceptibility to bias and overestimation of treatment impact, what might researchers, meta-analysts, and reviewers of the child and adolescent literature do to generate a more accurate picture of true population effects? The present findings suggest a few possible steps. First, reviewers in the child and adolescent area may need to follow the lead of those in other areas (see, e.g., Durlak, 1999; Durlak et al., 1991; Dush et al., 1989; Grossman & Hughes, 1992; Lipsey & Wilson, 1993; Saile, Burgeimer, & Schmidt, 1988; Shadish et al., 1993; M. L. Smith & Glass, 1977) and broaden their searches to include more unpublished research. Among the unpublished options, dissertations appear to offer special advantages. Because they are automatically entered in DAI, dissertations are free from publication–review bias and the file drawer effect compared with other nonjournal sources, such as invited presentations at professional meetings or manuscripts authors have not submitted. Other nonjournal sources are apt to lack the kind of quality control provided by dissertation committees; indeed, it may be the work of these committees that accounts for the remarkable methodological similarity we found between dissertation studies and published studies. Finally, we found, perhaps because of the absence of space constraints on dissertations, that they actually contain more of the information needed to understand study and treatment methods, to code studies, and to calculate ESs accurately (e.g., means and standard deviations of all outcome measures for all groups) than do published studies. Given the multiple strengths of dissertations, it seems important to include them in future meta-analytic efforts to assess the impact of child and adolescent treatment. One approach would be for reviewers and meta-analysts to routinely pool within their study samples all relevant published studies and dissertations that meet their methodological criteria and, in addition, to report findings separately for published studies and dissertations. This would provide a picture of the evidence base in the published literature and a comparison to the evidence base from methodologically acceptable research that is not subject to publication bias. The result might be a more accurate picture of the range within which true population ES values are likely to fall.

Given the importance of identifying the most representative sample of studies possible, it may also be worthwhile to consider steps that could be taken to identify unpublished studies other than dissertations. One could imagine, for example, a national registry of treatment outcome research, similar to that maintained through the International Cancer Research Data Bank (U.S. Dept. of Health and Human Services, 1983), intended to ensure as complete a documentation as possible of treatment outcome studies regardless of their outcome or publication status. The potential value of such a registry is illustrated by Simes’s (1987) finding that the pooled published literature showed a significant survival advantage for chemotherapy (over alkylating agent alone) in the treatment of advanced ovarian cancer but that registered trials showed no such advantage.

Returning to the question of what individual researchers can do to mitigate the effects of bias, our findings and previous literature suggest that sample size warrants attention. As our results show, the state of the art in child therapy research has been to use relatively small samples (about 16 per treatment group). This is a problem because publication bias is especially damaging to a field when the literature is characterized by numerous studies with small sample sizes (Begg, 1994; Begg & Berlin, 1988). Random variation ensures that ES estimates differ across studies, with the amount of variation determined, in part, by sample size. Large sample sizes generate ES estimates with low variation, whereas small sample sizes produce highly variable ES estimates (Begg & Berlin, 1988; Cooper & Hedges, 1994). As a result, aberrant ES estimates produced by chance are much farther from the true mean effect with small samples than with large samples (Begg, 1994; Begg & Berlin, 1988). Begg (1994) illustrated this effect through an example, which demonstrated that if five identical studies with small sample sizes of 20 per group each address the same topic, the largest mean ES produced by any one of these studies would be positively biased by 0.26 standard deviations. Studies using small sample sizes can therefore have a negative impact on a field because publication bias increases the likelihood that the positively biased results will be published. It is therefore evident that studies with small sample sizes tend to produce artificially inflated ES estimates that, because of publication bias, can create a literature characterized by ES overestimates (D. M. Lane & Dunlap, 1978). When the biasing effects of small sample studies and publication bias are present in a literature set, the mean ES estimates of the studies do not remain constant across changes in sample size. Indeed, in extreme cases, ESs are inversely related to sample size. In the present study, this inverse relation was observed in the published studies ($r = -.14, p < .05$), but not in the dissertations ($r = .01, p < .86$). This suggests that the inflated ES estimates in
the published studies result, in part, from the combined effects of publication bias and small sample studies.

The tendency of small sample sizes to produce highly variable ES estimates has led several authors within the medical field to argue that these studies should not be included in meta-analyses (Begg, 1994; Begg & Berlin, 1988; Berlin, Begg, & Louis, 1989; Easterbrook, 1991). They have argued that the use of small sample sizes has led to a collection of published works that have a greater number of statistically significant results than analysis of their power would indicate they should (Chase & Chase, 1976; Sedlmeier & Gigerenzer, 1989). Moreover, they have argued that only large, randomized clinical trials should be included within meta-analyses because if these studies overestimate ES, these deviations are far less than those with small sample studies. The fact that the mean ES of published studies was so much larger than the mean of dissertations in the present study is almost certainly due in part to the combined impact of small samples and publication bias.

Researchers certainly have the option of seeking larger samples, and reviewers and meta-analysts have the option of screening studies to exclude small samples. Our findings do not lead to an exact specification of ideal inclusion–exclusion rules, but we did find that with treatment group samples of 25 and lower, there was unusually wide dispersion in ES values and, thus, a large standard deviation (0.79). Standard deviations were markedly lower for studies with treatment group samples in the 26–50 range (0.48) and lower still in the 51 or more range (0.30). This is consistent with evidence from related fields that small samples produce high variability, which can exacerbate the impact of bias, and that large samples produce low variability that helps to mitigate the impact of bias (Begg & Berlin, 1988; Berlin et al., 1989; Easterbrook, 1991; Lipsey & Wilson, 1993).

Obtaining a large sample in clinical research is not always feasible or easily achieved. Moreover, as small sample studies constitute a majority of the studies in the child and adolescent psychotherapy literature, excluding these studies from meta-analyses may lead to an underestimation of the overall mean ES. Future meta-analysts, however, will need to take steps to address the published literature’s positive bias. They can do so by (a) using funnel graphs to check for the presence of publication bias (see Light & Pillemer, 1984) and (b) including dissertations so that the population’s true variability is more accurately represented.

Although our study offers useful findings, with implications for research and research synthesis, a few interpretive issues warrant attention. Although we assessed and controlled for multiple methodological and substantive differences between dissertations and published studies, it is possible that differences exist that we failed to detect and that some of these differences might account for some portion of the ES difference between the two study sets. As one example, elements of the therapeutic process (e.g., working alliance) might have systematically differed between the two study sets and contributed to the noted difference. Regrettably, few child psychotherapy studies measure therapy process (see Weersing & Weisz, 2002; Weisz, in press) making it impossible to test this important alternative hypothesis. As another example, graduate student’s ability to pilot, refine, and strengthen an intervention before testing is likely to be limited. It is possible that the greater research and treatment development experience, and the increased resources, typically associated with faculty and research center employment are associated with the development of more potent treatments than graduate students are likely to develop. Dissertations in this area tend to be each student’s first-ever clinical trial, whereas published studies have frequently been preceded by other outcome research done by the same investigators. Thus, it is certainly possible that some part of the larger ES we found for published studies reflects a genuine tendency for postdoctorate researchers to produce more potent treatments. If this is the case, then the “true population effect” of studies done after receipt of the doctorate may in fact be somewhat larger than the mean ES found for dissertations.

Overall, our findings highlight the need for (a) development and testing of treatments in randomized trials with samples large enough to generate reliable effects and (b) more comprehensive approaches to literature review and synthesis than have characterized the field thus far. These changes in emphasis could pay important dividends to the field, enriching research and sharpening our understanding of the true impact of child and adolescent treatment.

References

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


*Avila, A. L. (1985). The effect of directed imagery on the attention of...


*Conner, E. J. (1985). The relative effectiveness of two cognitive intervention approaches with attention deficit disordered children (Doctoral


Furlong, M. J. (1980). *Cognitive mediation training with hyperactive and nonhyperactive children: A metacognitive analysis* (Doctoral disserta-
tion, University of California, Santa Barbara, 1981). Dissertation Ab-
stracts International, 41, 0519.
training and covert positive reinforcement with elementary school chil-
behavior of aggressive boys (Doctoral dissertation, Texas A & M Uni-
*Grant, A. F. (1980). The effects of relaxation training on the test anxiety
of public school students in grades nine through twelve (Doctoral
dissertation, University of South Carolina, 1980). Dissertation Abstracts
International, 41, 990.
*Grant, J. E. (1987). A problem solving intervention for aggressive ado-
lescent males: A preliminary investigation (Doctoral dissertation, Syra-
*Grassi, R. (1984). Effects of self instructional training and rational emo-
tional education on emotional adjustment in elementary school children
(Doctoral dissertation, Hofstra University, 1984). Dissertation Abstracts
International, 46, 1730.
*Graybill, D., Jamison, M., & Swerdlik, M. E. (1984). Remediation of
impulsivity in learning disabled children by special education resource
teachers using verbal self-instruction. Psychology in the Schools, 21,
252–254.
ness in meta-analytic samples. Psychotherapy Research, 1, 17–24.
anxieties of elementary students. Psychology in the Schools, 25, 428–
436.
internalizing disorders: A review and analysis. School Psychology Re-
view, 21, 229–245.
training: Effects on social competence and problem behaviors (Doctoral
dissertation, Hofstra University, 1993). Dissertation Abstracts Interna-
tional, 53, 3477.
in adolescent offenders: 2. Intervention. Developmental Psychology, 26,
269–277.
modification techniques. Contemporary Educational Psychology, 12,
30–41.
intervention for elementary school children. Psychology in the Schools,
24, 161–164.
for parent–adolescent conflict. Social Work Research and Abstracts, 23,
3–8.
ioral parent training: Enhancement of treatment efficacy using a time-out
adolescents: A meta-analytic review. Journal of Clinical Child Psychol-
ogy, 27, 423–433.
Hansen, W. B., Collins, L. M., Malotte, C. K., Johnson, C. A., & Fielding,
Medicine, 8, 261–275.
Effects of a skills training intervention with juvenile delinquents. Re-
search on Social Work Practice, 1, 107–121.
Hedges, L. V. (1982). Estimation of effect size from a series of indepen-
(Eds.), The handbook of research synthesis (pp. 285–300). New York: Russell Sage Foundation.
Orlando, FL: Academic Press.
compliance training program for the treatment of noncompliance in
preschoolers (Doctoral dissertation, University of Kansas, 1985). Dis-
sertation Abstracts International, 45, 2689.
Attempting primary prevention. Canadian Journal of Counseling,
23, 273–287.
of child and adolescent group treatment: A meta-analytic review. Journal
of social reinforcement and time-out: Effects of brief parent training.
decrease impulsive behavior in severely disturbed/mentally retarded
subject (generalization, maintenance) (Doctoral dissertation, University
group assertive training on black adolescent aggression. Journal of
Consulting Psychology, 31, 95–98.
with aggressive male children in residential treatment (Doctoral disser-
tation, University of Northern Colorado, 1986). Dissertation Abstracts
International, 46, 2811.
for at-risk urban children (Doctoral dissertation, University of Connecti-
*Irvine, A. B., Biglan, A., Smolkowski, K., Metzler, C. W., & Ary, D. V.
(1999). The effectiveness of a parenting skills program for parents of
middle school students in small communities. Journal of Consulting &
Clinical Psychology, 67, 811–825.
*Israel, A. C., Stolmaker, L., & Andria, C. A. G. (1985). The effects of
training parents in general child management skills on a behavioral
psychotherapy: An integrated treatment for emotionally and behavior-
*James, M. R. (1990). Adolescent values clarification: A positive influence
on perceived locus of control. Journal of Alcohol and Drug Education,
*Jenson, J. M. (1988). Effects of a skills training intervention with juvenile
ing the quality of clinical trials for meta-analysis. JAMA, 282, 1054–1060.
*Jupp, J. J., & Griffiths, M. D. (1990). Self-concept changes in shy,
socially isolated adolescents following social skills training emphasizing
deficits: Analysis of components, generalization, and prediction of out-
come (Doctoral dissertation, Ohio University, 1984). Dissertation Ab-
stracts International, 44, 2895.
*Kaduson, H. G. (1993). Self-control game interventions for attention-
deficit hyperactivity disorder (Doctoral dissertation, Fairleigh Dickinson
gram for hyperaggressive children (Doctoral dissertation, Fordham Uni-


Received December 19, 2002
Revision received May 27, 2003
Accepted June 5, 2003