De Los Reyes and Kazdin (2009, Behavior Modification)

Manual for Range of Possible Changes Meta-Analysis

Andres De Los Reyes
University of Maryland at College Park

Alan E. Kazdin
Yale University

Outline

I. Brief Overview **pg. 3**

II. Techniques for Literature Search **pg. 4**

III. Gathering of Studies **pg. 6**

IV. The List of Studies We Need to Code **pg. 8**

V. Data Collection Part 1:
Coding Inclusion Criteria and Basic Study Characteristics **pg. 9**

VI. Data Collection Part 2: Coding Outcome Measures, Methods of Examination, and Statistical Analyses **pg. 16**

VII. Data Collection Part 3:
Coding Information to Calculate Effect Sizes **pg. 22**

VIII. Data Collection Part 4: Coding Classifications of Studies **pg. 25**

IX. Data Collection Part 5: Coding Classifications of Treatments **pg. 48**

Appendix-1. Sheet for Number of Studies **pg. 58**

Appendix-2. Operational definition of Cognitive-Behavioral Treatment for Childhood Anxiety Problems **pg. 60**

Appendix-3. Operational definition of Behavioral Parent Training for Childhood Conduct Problems **pg. 61**

Appendix-4. List of outcome measures that specifically assess childhood anxiety **pg. 62**

Appendix-5. List of outcome measures that specifically assess childhood conduct-related problems **pg. 63**

Appendix-6. Examples of Methods of Analysis and Types of Statistical Tests Employed in Treatment Outcome Studies **pg. 64**

Appendix-7. Methods of Calculating Effect Sizes and Statistical Significance Employed in this Meta-Analysis **pg. 66**

I. Brief Overview

Hello and thank you very much for participating in this research project. In this manual, you will learn how to gather studies for the project, how to code relevant characteristics of the studies, and how to classify the findings of studies, and treatments being studied. Because your initial experiences in this project are in assisting in the gathering of project data, it is important at this time that you not know about the hypotheses of the study. If you did know the study's hypotheses, the findings would not be useful, because one requirement of meta-analyses is that data collection be done by people who do not know what the expected findings are. Thus, if you were aware of the study's hypotheses, the results of the study would not be publishable. At the same time, a key aspect of this project will be full disclosure to you of both study hypotheses and results of the study, once the study is complete.

Below is my contact information. If you have any questions over the course of the project, please feel free to contact me. Any questions you may have, please err on the side of caution, and contact me with your concern.

Contact information:

Andres De Los Reyes, M.S., M.Phil.
Email: andres.delosreyes@yale.edu
Office: Kirtland B01A
Mobile: (305) 905-5149

II. Techniques for Literature Search

The literature search will be conducted by year. Fortunately, we only have to search for 4 years ourselves (2003, 2004, 2005, 2006), because we have a list of all relevant studies up until 2002. A single person will be responsible for compiling the list of studies to go through for any given year. The two methods for searching are listed below, and they must be done exactly this way, in order to remain consistent with how other researchers have done these kinds of reviews.

The first thing you will do is conduct searches using two search engines: PsychInfo and PubMed (in PubMed, you will only be searching the MEDLINE database; there are instructions for doing this below). After conducting the searches, please compile the list of studies (see instructions below in this section).

Terms to use for PsychInfo

Term box 1: client-centered OR contract- OR counseling OR cotherapy OR dream analysis OR insight- OR intervention- OR model- OR modifica- OR operant- OR paradox- OR psychoanaly- OR psychodrama- OR psychothera- OR reinforce- OR respondent OR role-playing OR therap- OR training OR transactional OR treatment

> * Please limit this box to "human," "English language," and the year you are searching (either 2003, 2004, 2005, or 2006)

Term box 2: adolescen- OR child- OR juvenile- OR pre-adolescen- OR youth-

> * Please limit this box to "human," "English language," and the year you are searching (either 2003, 2004, 2005, or 2006)

Term box 3: assess- OR comparison OR effect- OR efficacy OR evaluat- OR influence OR impact OR outcome

> * Please limit this box to "human," "English language," and the year you are searching (either 2003, 2004, 2005, or 2006)

Combining term boxes: Once the three term boxes are composed, please write down the number of studies for each term box (sheet is in Appendix-1), after all limits have been imposed (e.g., row that reads "limit X to [human and English language and yr=….]"). Then, combine the three term boxes using the (AND) term. To do this, just click on "Combine Searches" and click on the relevant term boxes. Please write down the final study count for the search on the sheet in Appendix-1.

Terms to use for MEDLINE via PubMed

Search "PubMed" for "mental disorders"

Click on "Limits" then go to "Limited to" and please do the following:

-Leave "All fields" as is, and leave "only items with abstracts" unchecked
-Choose "Clinical Trial"
-Choose "English" and choose "MEDLINE"
-Choose "All Child: 0-18 years" and choose "Humans"
-Please leave "Gender" as is
-Please leave "Entrez Date" as is, and please leave "Publication Date" as is
-For "From," please type the year you are searching and leave the other two boxes blank, and for "To," please type the year you are searching and leave the other two boxes blank
-Finally, please record the number of studies that came up in your search for your year in the Appendix-1 sheet

Compiling the Lists of Studies for PsychInfo

-After you have completed your search and the list is up, go to the “Results Manager” and make sure that “Citation + Abstract” is selected under “Fields,” and click “Display.”

Compiling the Lists of Studies for MEDLINE via PubMed

-After you have completed your search and the list is up, below the tab “Limits” on the top of the page, you will see “Display.” To the right of “Display,” please click “Abstract.”

III. Gathering of Studies

Ok, so we now have the list of studies for the three years we have searched. It turns out that we already have most of the list compiled from another study! Thus, what we have to do is add to this list our own list of studies from 2003, 2004, 2005, and 2006, so that it is as current as possible. However, in order to do that, we have to dwindle down our huge lists of studies from the four years to studies that meet criteria to make it on the list that has already been compiled from years past.

Please apply the following criteria for the studies from the lists:

(1) Study must be peer-reviewed. What this usually means is that the study was published in a peer-reviewed journal. Now, PsychInfo citations usually have this information in the citation. However, MEDLINE via PubMed does not. Basically, if the name of the place the study was published follows a similar citation format as the following, "Journal Name. 2004; 99(2): 99-9," then you should be confident the study was published in a peer-reviewed journal. Please exclude from your lists unpublished dissertations that come up in the search, as well as unpublished manuscripts, and book chapters. The one exception may be if it looks like the book chapter was peer-reviewed, but this would be rare.

(2) Studies must be tests of psychotherapy. What we mean here is a test of any intervention designed to alleviate non-normative psychological distress, reduce maladaptive behavior, or increase deficient adaptive behavior through counseling, interaction, a training program, or a predetermined treatment plan.

(3) Studies must also include the following: **(a)** include a comparison of psychotherapy to a control group (waitlist, no treatment, pill placebo, or other process intended to be inert); **(b)** involve prospective design (outcome measures to see if the treatment worked must be administered before and after treatment) and random assignment of subjects to treatment and comparison conditions; **(c)** use a sample within the 3- to 18-year-old age range; **(d)** use participants selected for having psychological problems or maladaptive behavior; and **(e)** include a post-treatment assessment of each psychological problem or maladaptive behavior for which participants were selected and treated.

(4) A critical factor will be weeding out those studies where there is any effect of medication on any participants in the studies, in order to remain consistent with prior reviews. To ensure that the focus of the articles was on comparing psychotherapy to a comparison group, please include only those studies in which participants in the groups being compared were not taking any psychotropic medications, or medications being taken to alleviate or treat psychological or psychiatric problems. What this means is that any article that explicitly states that participants in the conditions being compared were taking medication should not be included on the study list.

(5) Another key factor will be weeding out studies where no significant findings of the intervention were found on measures of the problem targeted in the intervention. To ensure this, of the studies identified with the above criteria, please only include studies that reveal a statistically significant benefit of the intervention examined, relative to a control condition on at least one outcome measure assessing the construct or behavior that was the focus of the treatment. Please consult Appendices 4 and 5 to determine which measures may be used to make decisions on this criterion.

(6) The last methodological factor that will be used to include and exclude studies will be that the study examined treatment outcomes on at least three measures of the construct or behavior that was the focus of the intervention. Each study must use at least three measures (questionnaires, structured interviews, independent assessments [like laboratory or home-based observations]) of the primary target of the intervention. As you will see below, this either means that treatments of child anxiety include at least three outcome measures focused specifically on anxiety, and treatments of child conduct include at least three outcome measures focused specifically on conduct. Please consult Appendices 4 and 5 to make decisions regarding this criterion.

(7) Great, so now that you have gotten this far, there are two more criteria that we need to employ. First, of the studies that meet ALL the criteria above, the studies that we are to include on the list must be those that do the two following things:

(a) Examine a psychological treatment for EITHER of the following problems:

> **(1)** Anxiety-related problems and disorders. These include any of the phobias (e.g., simple or specific phobias or fears of animals or blood or natural disasters or anything else, social phobia), or anxiety disorders or anxious avoidance (e.g., generalized anxiety disorder, separation anxiety disorder, post-traumatic stress disorder, overanxious disorder, test anxiety, public speaking anxiety, shyness).
>
> **(2)** Conduct-related problems and disorders (e.g., oppositional defiant disorder, conduct disorder). To be repetitive, these include any problems with aggression, anger, conduct, delinquency, fire-setting, basically anything conduct-related.

(b) Of those treatments for the problems noted above, those treatments that can be described as the following:

> **(1)** For anxiety-related problems and disorders, those treatments that can be described as "youth-focused cognitive-behavioral therapy" (see Appendix-2). The treatment must be stand-alone, and not combined with a non-youth-focused treatment component (e.g., parent treatment sessions), or a youth-focused non-CBT component (e.g., child relationship therapy).
>
> **(2)** For conduct-related problems and disorders, those treatments that can be described as "parent-focused behavioral parent training" (see Appendix-3). The treatment must be stand-alone, and not combined with a non-parent-focused treatment component (e.g., child treatment sessions), or a parent-focused non-BPT component (e.g., parent-child interaction therapy).

Now, the list of treatment studies we have so far includes 8 studies of CBT for anxiety, and 5 studies of behavioral parent training for conduct problems. So, when doing this search for a given year, please do not be surprised if your list of treatments is quite short. We would be surprised if, across the three years, we found more than 5 or so studies for either treatment.

IV. The List of Studies We Need to Code

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

(2) Flannery-Schroeder, E.C., & Kendall, P.C. (2000). Group and individual cognitive-behavioral treatments for youth with anxiety disorders: A randomized clinical trial. *Cognitive Therapy and Research, 24,* 251-278.

(3) Gallagher, H.M., Rabian, B.A., & McCloskey, M.S. (2004). A brief group cognitive-behavioral intervention for social phobia in childhood. *Journal of Anxiety Disorders, 18,* 459-479.

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

(5) Kendall, P.C., Flannery-Schroeder, E., Panichelli-Mindel, S.M., Southam-Gerow, M., Henin, A., & Warman, M. (1997). Therapy for youths with anxiety disorders: A second randomized clincal trial. *Journal of Consulting and Clinical Psychology, 65,* 366-380.

(6) King, N.J., Tonge, B.J., Mullen, P., Myerson, N., Heyne, D., Rollings, S., et al. (2000). Treating sexually abused children with posttraumatic stress symptoms: A randomized clinical trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 1347-1355.

(7) Leal, L.L., Baxter, E.G., Martin, J., & Marx, R.W. (1981). Cognitive modification and systematic desensitization with test anxious high school students. *Journal of Counseling Psychology, 28,* 525-528.

(8) Leung, C., Sanders, M.R., Leung, S., Mak, R., & Lau, J. (2003). An outcome evaluation of the implementation of the Triple P-Positive Parenting Program in Hong Kong. *Family Process, 42,* 531-544.

(9) McMurray, N.E., Bell, R.J., Fusillo, A.D., Morgan, M., & Wright, F.A.C. (1986). Relationship between locus of control and effects of coping strategies on dental stress in children. *Child & Family Behavior Therapy, 8,* 1-17.

(10) Spence, S.H., Donovan, C., & Brechman-Toussaint, M. (2000). The treatment of childhood social phobia: The effectiveness of a social skills training-based, coginitive-behavioral intervention, with and without parental involvement. *Journal of Child Psychology and Psychiatry, 41,* 713-726.

(11) Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52,* 666-678.

(12) Webster-Stratton, C. (1990). Enhancing the effectiveness of self-administered videotape parent training for families with conduct-problem children. *Journal of Abnormal Child Psychology, 18,* 479-492.

(13) Webster-Stratton, C. (1992). Individually administered videotape parent training: ‘‘Who benefits?’’ *Cognitive Therapy and Research, 1992, 16,* 31-52.

(14) Webster-Stratton, C., Kolpacoff, M., & Hollinsworth, T. (1988). Self-administered videotape therapy for families with conduct-problem children: Comparison with two cost-effective treatments and a control group. *Journal of Consulting and Clinical Psychology, 56,* 558-566.

(15) Webster-Stratton, C., Reid, M.J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology, 33,* 105-124.

(16) Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology, 65,* 93-109.

V. Data Collection Part 1: Coding Inclusion Criteria and Basic Study Characteristics

The first thing we will do is code some of the descriptive characteristics of the studies, as well as the characteristics of the studies that warranted their inclusion in the study. You can find most of these characteristics in the Method section of the articles. Do not let the title of this section fool you: Some of these characteristics are tough to find in some of these articles, so please double-check all of your work, and please call or email me with any questions that you may have.

NOTE. WHEN MAKING CODES IN SHEET, PLEASE CODE MISSING DATA USING THE CODE "888." PLEASE CODE "999" FOR ITEMS THAT ARE NOT APPLICABLE TO THE STUDY YOU ARE CODING.

Inclusion Criteria

(1) *Study must have been published in a peer-reviewed journal*. You can usually find this out from a combination of two sources. The first source is usually listed somewhere on the first page, with a volume number next to it, and page numbers; this usually signifies that the source of the article is a peer-reviewed journal. Second, somewhere on the first page or the last page of the article will be dates that the article was first submitted for publication, as well as the date that the article was accepted for publication. This usually signifies that the article underwent some sort of review process before it was published. If you find this information in the article, then you should be confident that the study reported in the article was published in a peer-reviewed journal.

(2) *Studies must be tests of psychotherapy.* If the study reported in the article tested at least one form of psychotherapy, then it is included in the meta-analysis.

(3) *Study must include comparison of the treatment examined with a control group.* Study must report that the treatment(s) they examined was/were compared to a control group. The authors will usually clue you into this by including a comparison group that they described as: **(a)** waitlist; **(b)** no treatment; **(c)** pill placebo; or **(d)** some other process intended to be completely inert, like an attention-placebo or psychological placebo (e.g., meeting with therapist weekly and talking about the weather, no directive therapeutic service).

(4) *Study must involve prospective design.* What this means is that outcome measures were administered to see if the treatment worked, and these outcome measures were administered both before and after treatment.

(5) *Study must employ random assignment of subjects to treatment and comparison conditions.* The author states this somewhere in the Abstract or Method.

(6) *Study must employ a sample within the 3- to 18-year-old age range.* You can usually find the author explicitly stating the age range of the sample somewhere in the Abstract or Method.

(7) *Study must examine participants selected for having psychological problems or maladaptive behavior.* The children being treated for problems must be selected for having those problems. What this means is that the children were either assessed for or diagnosed with the problems targeted for treatment, or referred to a clinic for the problems targeted for treatment, or parents brought their children in for treatment for the problems targeted for treatment. See Appendices 4 and 5 for lists of outcome measures for the problems targeted in the treatments examined.

(8) *Study must include a post-treatment assessment of each psychological problem or maladaptive behavior for which participants were selected and treated*. What this means is that if the children in the sample were treated for a specific problem (e.g., child anxiety), measures of that problem were administered after the treatment to see if the treatment changed that problem. See Appendices 4 and 5 for lists of outcome measures for the problems targeted in the treatments examined.

(9) *Examine a psychological treatment for EITHER of the following problems:*

- Anxiety-related problems and disorders. These include any of the phobias (e.g., simple or specific phobias or fears of animals or blood or natural disasters or anything else, social phobia), or anxiety disorders or anxious avoidance (e.g., generalized anxiety disorder, separation anxiety disorder, post-traumatic stress disorder, overanxious disorder, test anxiety, public speaking anxiety, shyness).

- Conduct-related problems and disorders (e.g., oppositional defiant disorder, conduct disorder). To be repetitive, these include any problems with aggression, anger, conduct, defiance, delinquency, fire-setting, basically anything conduct-related.

(10) *Of those treatments for the problems noted above, those treatments that can be described as the following:*

- For anxiety-related problems and disorders, those treatments that can be described as "youth-focused cognitive-behavioral therapy" (CBT) (see Appendix-2). The treatment must be stand-alone, and not combined with a non-youth-focused treatment component (e.g., parent treatment sessions), or a youth-focused non-CBT component (e.g., child relationship therapy).

- For conduct-related problems and disorders, those treatments that can be described as "parent-focused behavioral parent training" (BPT) (see Appendix-3). The treatment must be stand-alone, and not combined with a non-parent-focused treatment component (e.g., child treatment sessions), or a parent-focused non-BPT component (e.g., parent-child interaction therapy).

- **Critical note:** For many of the studies in the meta-analysis, often multiple treatments are examined at the same time. For a study to be included in the meta-analysis, it only has to examine 1 treatment that meets the criteria above. Also, that 1 treatment has to be a stand-alone treatment that can meet either of the descriptions in Appendices 2 or 3. Stated another way, the treatment cannot be combined with another treatment or an add-on component (e.g., youth-focused cognitive-behavioral therapy plus family therapy, where parents are involved in the treatment also). **However,** for CBT for child anxiety, it is alright to include a treatment if the authors mention that the parent was involved in a session or two over the course of the child's treatment for information-gathering purposes (i.e., no therapeutic tools or benefits were administered to the parent).

(11) *Participants in the groups being compared were not taking any psychotropic medications, or medications being taken to alleviate or treat psychological or psychiatric problems.* Any article that states that participants in the conditions being compared were taking medication should not be included on the study list. This does not mean that there cannot be a medication-only condition in the study; just that the groups being compared in the meta-analysis cannot have participants in them that are taking medication.

(12) *Study must examine treatment outcomes on at least three measures of the construct or behavior that was the focus of the intervention.* What this means is that if the treatment examined was supposed to target child anxiety, then at least three outcome measures that specifically assess anxiety must have been administered. The same for child conduct problems: at least three outcome measures that specifically assess conduct problems must have been administered. Stated another way, after treatment finished three outcome measures assessed how the participants in the sample were doing, with regard to the specific problem they were treated for. See Appendices 4 and 5 for lists of outcome measures for the problems targeted in the treatments examined.

(13) *Study must reveal a statistically significant benefit of the intervention examined, relative to a control condition.* The treatment examined must be statistically significantly different from a control condition (one of the conditions noted in Item 3 of this section) on at least one measure of the problem or behavior targeted for treatment. For CBT, the intervention examined must be significantly better than a control condition on at least one anxiety measure. For BPT, the intervention examined must be significantly better than a control condition on at least one conduct measure. Note that for studies of multiple interventions, this criterion may be met for some interventions and not others. Thus, please examine this criterion for each treatment that has met criteria for inclusion in the study thus far.

Note that a study may meet this criterion if at least one intervention in the study satisfies this criterion (even if others do not). See Appendices 4 and 5 for lists of outcome measures for the problems targeted in the treatments examined.

Basic Study Characteristics

(1) *Author name*. Please provide the last name of the author listed first in the article. You can find this name on the front page of the article.

(2) *Year of publication*. Please provide the year in which the article was published. You can find this information on the front page of the article, usually on the very top of the page.

(3) *Treatment type and primary construct targeted for treatment*. Please code whether the study examined:

(1) youth-focused cognitive-behavioral treatment(s) for child anxiety problems
(2) parent-focused behavioral parenting treatment(s) for child conduct problems

(4) *Type and number of control group(s)*. Please code the number and type(s) of control group(s) used in the study:

(1) waitlist
(2) no treatment
(3) pill placebo
(4) an attention-placebo or psychological placebo or otherwise inert process (i.e., a group that the authors expected to exhibit no significant change in the target problem over the time that the treatment(s) examined in the study is administered to other groups).

You can find this information in the Abstract or the Method section. Please note that a given article may have more than one control group that meets for the above criteria. If the article notes that more than one inert control group was used in the study, please list all control groups here.

(5) *Type and number of treatment group(s) (e.g., individual, group)*. Of those treatments examined in the study that meet criteria for inclusion in the meta-analysis, please code the type of treatment group(s) examined in the study as:

(1) individual
(2) group

Please note that a given article may examine multiple treatments within a given study. At the same time, even though multiple treatments may be examined, not all of them may meet criteria for examination in this meta-analysis. For instance, in a study examining youth-focused cognitive behavioral treatment for child anxiety, a treatment may be examined that combines the youth-focused treatment with a treatment focused on the parents. This combined treatment would not be included in the coding, and thus not examined in this meta-analysis. Only a stand-alone youth-focused treatment would be eligible for inclusion in the study. Therefore, please code multiple treatments for type of treatment if more than one treatment is examined in the study, but only code those treatments that meet criteria for inclusion in the meta-analysis. Please consult Appendices 2 and 3 when coding this information.

(6) *Sample type*. Please code the type of sample treated in the study:

(1) diagnosed referred/recruited outpatients
(2) symptomatic referred/recruited outpatients
(3) diagnosed school sample
(4) symptomatic school sample

The distinction between sample types boils down to the setting in which treatment occurs (clinic or university clinic versus school), and how the children were assessed prior to treatment (diagnosed with a disorder that will be targeted in treatment or assessed with a clinical measure that identifies problematic children prior to treatment, usually with some sort of cutoff score). You can usually find this information in the Abstract, or the very beginning of the Method section (often in the Participants subsection of the Method).

(7) *Number of participants in each treatment and control group identified in Items 4 and 5* ***(at baseline)***. Please code the number of children in the treatment and control groups identified in Items 4 and 5 of this section **at baseline** (i.e., number of participants at the beginning of treatment, before any dropouts or attrition occurred). You can usually find this information in a table in the article's Method section describing the characteristics of the groups. Otherwise, this information is reported in text in the Method section, often in the Participants subsection of the Method. **Note on sample size and BPT studies: For BPT for child conduct problem studies, getting the sample sizes for children in conditions may take some inferring. For instance, some studies may only tell you the number of parents in each treatment and control condition, and not the number of children; you may have to infer number of children from number of parents. One way you could do this is to peek into the characteristics of parents. For example, the study may mention that 20 mothers and 15 fathers were involved in the control condition, but then also mentions that 75% of the families in the condition were married and the only single parents in the control condition were all mothers. Given this information, you can infer that there were 20 children in the control condition, because all 15 fathers and 15 of the mothers were married to each other, leaving the 5 single mothers, and all totaling 20 children. Of course, if you have to infer sample size like this, please make sure that the resulting number for each condition matches up to the total sample size coded in Item 30 below.**

(8, 8a.) *Length of treatment for each treatment group being examined*. Please code the duration(s) of the treatment(s) being examined in the study, up until post-treatment assessment. In other words, codes in this item should not include booster sessions (sessions conducted after treatment is complete) in the total length of treatment, if the booster sessions occurred after the post-treatment assessment upon which treatment conditions were compared. Treatment duration will be converted into minutes. For instance, if the article states that the treatment group met for 8 1-hour sessions, treatment duration should be coded as 480 minutes. If treatment sessions are reported in the text as a mean (e.g., children were seen for a mean of 10 sessions [12-14]), please use the mean. Similarly, if a range of treatment lengths is given in the text (50-60 minutes), please take the mean of this range to compute treatment length. Treatment length will be coded twice. First, please code the length of treatment the authors manualized (i.e., the length of treatment the authors codified in manual form that all participants *should* have received). Second, sometimes the authors report treatment length in terms of the average treatment that was *actually* received by participants in the study; if this information is available, please code this treatment length as well. **Note. Sometimes in BPT studies for child conduct problems, treatment length will be measured separately for male and female parents. In those cases, please use the following formula for combining treatment length across parents:**

$$\textbf{Length (min.)} = [\,(N_{Mothers})(Length_{Mothers}) + (N_{Fathers})(Length_{Fathers})\,] \;/\; (N_{Mothers} + N_{Fathers})$$

(9) *Percentage of participant dropout in sample.* Please code the percentage of participants in the total study sample that dropped out of their respective experimental condition before the end of the study (i.e., before they could be assessed for problem behavior after treatment ended). This should include treatment of all participants, regardless of whether they were in treatment or control groups, or in a treatment or control group that was not identified in Items 4 and 5 of this section (i.e., total sample attrition). You can usually find this information in the Method section, or in the beginning of the result section.

(9a) *Percentage and number of participant dropout for each treatment and control group identified in Items 4 and 5.* Please code the percentage and number of participants in each treatment and control condition that dropped out of their respective experimental condition before the end of the study (i.e., before they could be assessed for problem behavior after treatment ended). This should include treatment of all participants that were identified in Items 4 and 5 of this section. You can usually find this information in the Method section, or in the beginning of the result section.

(10) *Age range of children in sample (in years).* Please code the age range of the children in the sample (e.g., children in the sample were 3 to 5 years of age). You can usually find this information in the Abstract, or the very beginning of the Method section (often in the Participants subsection of the Method).

(11) *Mean age of children in total sample (in years).* Please code the mean age of the children in the total sample (e.g., the mean age of children in the sample was 5.3 years of age). You can usually find this information in the Abstract, or the very beginning of the Method section (often in the Participants subsection of the Method). **Note on age in months. Some authors only report age in months for the sample and for children in each experimental condition. If age is reported in months only, please convert to years by dividing the number of months by 12. If age is reported by years and months, please convert the months to a decimal of years by dividing the months by 12 and then adding it to the years figure.**

(12) *Mean age of children in each experimental condition (in years)***.** Please code the mean age of the children in each treatment and control condition identified in Items 4 and 5 (e.g., the mean age children in the control condition was 5.3 years of age). You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method), or in a table in the article's Method section describing the characteristics of the treatment and control groups. **Note on age in months. Some authors only report age in months for the sample and for children in each experimental condition. If age is reported in months only, please convert to years by dividing the number of months by 12. If age is reported by years and months, please convert the months to a decimal of years by dividing the months by 12 and then adding it to the years figure.**

(13) *Gender of children in sample.* Please code the percentage of boys in the total sample. You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method).

(14) *Gender of parent (for BPT).* For studies of BPT for child conduct problems, please code the percentage of female parents in the total sample. You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method). **Note on coding BPT parent gender: This may also be a tricky code to calculate, because sometimes BOTH mother and father are involved in treatment for the same child. Please code this item in terms of the percentage of cases that a female (usually mother) is involved in treatment.**

(15) *Percentage cases male parent involved in treatment (for BPT).* For studies of BPT for child conduct problems, please code the percentage of male parents in the total sample. You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method). **Note on coding BPT parent gender: This may also be a tricky code to calculate, because sometimes BOTH mother and father are involved in treatment for the same child. Please code this item in terms of the percentage of cases that a male (usually father) is involved in treatment.**

(16) *Percentage cases both parents involved in treatment (for BPT).* For studies of BPT for child conduct problems, please code the percentage of both parents being involved in treatment in the total sample. You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method). **Note on coding BPT parent gender: This may also be a tricky code to calculate, because sometimes BOTH mother and father are involved in treatment for the same child, and sometimes only one parent is involved. Please code this item in terms of the percentage of cases that both parents are involved in treatment.**

(17) *Age range of parents in sample (for BPT).* For studies of BPT for child conduct problems, please code the age range of the parents in the total sample (e.g., parents in the sample were 34 to 55 years of age). You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method). **Note on calculating age range: Sometimes age statistics are reported for mothers and fathers separately in the sample, so please make sure to consolidate ranges across parents to reach the coded information for this item.**

(18) *Mean age of parents in sample (for BPT).* For studies of BPT for child conduct problems, please code the mean age of the parents in the total sample (e.g., the mean age of parents in the sample was 36.5 years of age). You can usually find this information in the very beginning of the Method section (often in the Participants subsection of the Method). **Note on calculating mean age: Again, sometimes age statistics are reported for mothers and fathers separately in the sample, so please make sure to consolidate ranges across parents to reach the coded information for this item. Please use this formula to calculate ages reported for different genders:**

$$\textbf{Mean age} = [\,(N_{\text{Mothers}})(\text{Mean age}_{\text{Mothers}}) + (N_{\text{Fathers}})(\text{Mean age}_{\text{Fathers}})\,] \;/\; (N_{\text{Mothers}} + N_{\text{Fathers}})$$

(19) *Were children clinic-referred or recruited?* Please code whether (referred = **1**) or not (recruited = **2**) children targeted for treatment in the study were referred to the clinic (e.g., directed or suggested by school counselors, psychiatric institution, or other health professionals [or self-referred] to undergo treatment at the setting in which the treatment(s) examined was/were administered), or recruited through advertisements or announcements. This information can usually be found in the Method section (often in the Participant subsection of the Method).

(20) *Therapist type*. Please code whether the majority of therapists in the study were clinicians by vocation (code this **1**), or whether the majority of therapists in the study were not primarily clinicians by vocation; this latter group includes graduate students, researchers, and/or professors (code this **2**). This information can usually be found in the Method section, typically in a Treatment subsection of the Method that explains the structure and characteristics of treatment or treatments examined in the study.

(21) *Treatment setting*. Please code whether the study examined a treatment or treatments in a clinical service setting, such as an outpatient hospital program (code this **1**), or a nonclinical setting, such as a university clinic, lab clinic, or primary or secondary school (code this **2**). This information can usually be found in the Method, typically in a Treatment subsection of the Method that explains the structure and characteristics of treatment or treatments examined in the study, or in the Participant subsection of the Method.

(22) *Sample diagnosed/identified as problematic before treatment?* Please code whether children in the sample were diagnosed for the problem they were going to be treated for (code this **1**), whether they were assessed using cutoff scores derived from ratings from a measure that assesses the problem they were going to be treated for (code this **2**), a combination of **1** and **2** (identified as problematic before treatment using both cutoff scores and diagnoses; code this **3**), or were just assessed prior to treatment (code this **4**). This information is usually found in the Method, and you will usually know how to code this information when the authors describe how they included and excluded participants into their study. If the authors note that children had to meet diagnostic criteria for a disorder to be included in the study, that usually means they assessed and diagnosed the children. Also, if the authors described a structured interview used to diagnose the children prior to treatment, this usually means they assessed and diagnosed the children. The same is true for those children that were identified as problematic using cutoff scores derived from ratings from a measure.

(23) *If diagnosed sample, what were they diagnosed with?* Please code the instrument or interview used to diagnose the sample as problematic before treatment. A list of diagnostic interviews can be found in the coding sheet. This information is usually found in the Method, typically in the Measures subsection of the Method.

(24) *If diagnosed sample, what system were they diagnosed under?* Please code the system used to diagnose the sample as problematic before treatment. A list of diagnostic systems can be found in the coding sheet. This information is usually found in the Method section, typically in the Measures subsection of the Method, or the Participants subsection of the Method.

(25) *If assessed using cutoff scores, what measure(s) was/were used?* Please code the instrument used to identify children as problematic before treatment, if they were identified using a measure, and some sort of cutoff score on that measure was used by the authors to identify problematic youths to treat. A list of such measures can be found in the coding sheet. This information is usually found in the Method, typically in the Measures subsection of the Method, or the Participants subsection of the Method.

(26) *If assessed using cutoff scores, what type of cutoff score was used?* Please code the type of cutoff score the authors used to identify problematic children to treat, if they were identified using a measure, and some sort of cutoff score on that measure was used by the authors to identify problematic youths to treat. A list of the types of cutoff scores can be found in the coding sheet. This information is usually found in the Method, typically in the Measures subsection of the Method, or the Participants subsection of the Method.

(27) *Total number of control conditions studied, regardless of study inclusion.* Please code the total number of control conditions the authors examined in the study, including those control conditions the authors examined that did not meet criteria for inclusion in the meta-analysis. Essentially, Item 27 is the number of groups coded in Item

4, **PLUS** all other control conditions studied (i.e., all conditions in the study to which the authors were interested in comparing their treatments of interest).

(28) *Total number of treatment conditions studied, regardless of study inclusion.* Please code the total number of treatment conditions the authors examined in the study, including those treatment groups the authors examined that did not meet criteria for inclusion in the meta-analysis. Essentially, Item 28 is the number of groups coded in Item 5, **PLUS** all other treatment conditions studied.

(29) *Total number of children in entire study, regardless of study inclusion.* Please code the total number of children examined in the entire study, including those children that were in treatment and control conditions that were not included in the meta-analysis. You can usually find this total in the Abstract of the article, or at the very beginning of the Method, in a Participants subsection of the Method.

(30) *Total number of children in Item 7 (all treatment and control groups combined)***.** Finally, please code the total number of children examined in the treatment and control conditions identified for inclusion in the meta-analysis. You can usually find this total in the Abstract of the article, or at the very beginning of the Method, in a Participants subsection of the Method.

VI. Data Collection Part 2: Coding Outcome Measures, Methods of Examination, and Statistical Analyses

In this section, we are interested in coding: (1) characteristics of and findings based on measures of outcome; and (2) characteristics of and findings based on the results of statistical analyses. The coding sheet is set up to accommodate coding, but again, if you have any questions please feel free to contact me via email or phone.

(1) *Name of outcome measure.* Please code the name/abbreviation of the measure.

(2) *Type of outcome measure methodology*. For each outcome measure, please code whether the measure was taken from a:

(1) questionnaire/report/rating scale format
(2) behavioral lab or home observation (i.e., usually by an independent observer)
(3) structured interview (usually employed to arrive at diagnoses of disorders, symptom counts, or measures of disorder severity)
(4) composite score (multiple sources/measure methodologies)

You can usually find this information in the Method section, often in a Measures subsection of the Method.

(3) *Information source that completed outcome measure*. For each outcome measure, please code whether the measure was completed by any of the following:

(1) child self-report
(2) parent report, **Mother**
(3) parent report, **Father**
(4) parent report, **Unspecified (sometimes mother or father, depending on child)**
(5) teacher report
(6) independent observer (laboratory or home)
(7) clinical interviewer based on child report
(8) clinical interviewer based on parent report, **Mother**
(9) clinical interviewer based on parent report, **Father**
(10) clinical interviewer based on parent report, **Unspecified (sometimes mother or father, depending on child)**
(11) clinical interviewer based on teacher report
(12) clinical interviewer based on multiple reports (e.g., parent/child; parent/teacher)
(13) composite report (multiple sources in one measure)

Please keep in mind that in a given study, multiple outcome measures can be coded for information sources that fall under categories “7” through “11,” because the same structured interview may be administered to multiple people in a study. However, not all of these measures may have information about them in the Results, as far as whether they were actually used after the intervention to assess outcomes. For example, in a child anxiety intervention study, a structured interview may have been given by a clinician to both parent and child before treatment, but outcomes on that structured interview may only be provided after treatment based on clinicians’ interviews of parents. Thus, as in all items in this section, please only code those information sources for outcome measures that were employed in the study to assess outcomes after the intervention was complete. **For items in which information source needs to be coded, please only code a given outcome measure using the single code that best represents it (e.g., a clinician interview measure based on mother’s report should only be coded “8,” and NOT “2” and “8”).**

NOTE. FOR ITEMS 4-7 BELOW, PLEASE ALIGN OUTCOME FINDINGS ACROSS ITEMS ON THE CODING SHEET, SO THAT YOU ARE PROVIDING INFORMATION FOR THE SAME OUTCOME FINDING ACROSS ITEMS (E.G., INFORMATION FOR A FINDING CODED IN 4A IS CODED IN 5A, 6A, AND 7A).

(4) *Method(s) of analysis/analyses employed to examine outcome measure*. For each outcome measure, please code whether the measure was examined using any of the following outcome methodologies:

(1) tests of mean differences
(2) tests of diagnostic status
(3) tests of clinically significant change

Please keep in mind that a given outcome measure may be examined using multiple methodologies. For example, an outcome measure may be used to examine between-condition mean differences and clinically significant change. For these instances, a single outcome measure may be examined using two different methods if the same score from the measure is used for each method. An example may be a self-report anxiety measure's total score that is examined using mean differences comparisons, as well as tests of clinically significant change. In this example, the same usage of the measure (total score) is kept constant between the two methods of examining outcome, and should be coded within the same outcome number (e.g., mean differences: Outcome 1, 4a; clinically significant change: Outcome 1, 4b). If the same measure was examined using different data using two different methods (e.g., total score for mean differences, subscale score for clinically significant change), these two ways of gauging change should be treated as two different outcome measures (e.g., total score for mean differences: Outcome 1, 4a; subscale score for clinically significant change: Outcome 2, 4a). Thus, the categories are not mutually exclusive, and articles should be examined closely to see if a given outcome measure is examined using more than one method. Examples of the different methods of analysis are illustrated in Appendix-6. You can usually find this information in the Results section.

Note. Very important, tests of diagnostic status are their own method of analysis, and do not count as a test of clinically significant change, even if the authors identify tests of diagnostic status as a test of clinically significant change.

(5) *Type(s) of statistical test/tests employed to examine outcome measure*. For each outcome measure, please code the type of statistical test that was or will be employed to examine outcomes. This code should be applied within each individual outcome measure finding. Again, please keep in mind that types of statistical tests may not be applied to all outcome measures. For instance, when treatment and control conditions are compared on diagnostic status, this test may only be examined using chi square, whereas tests of mean differences between conditions may be examined using *t* tests. Please also note that a given kind of outcome measure (e.g., questionnaire measure of child anxiety symptoms) may be examined using multiple kinds of methods. Stated another way, just because an outcome measure is examined using one test does not mean it is not examined using another test later on in the article. The tests that you may see fall into two categories:

(1) *t* test
(2) chi square
(3) Other statistical test (please describe)

A helpful note with Item 5 is that you can usually find information to calculate mean differences by looking at the Tables in the article. However, tests of diagnostic status and clinically significant change are usually reported in the text of the article. Examples of the different types of statistical analysis are illustrated in Appendix-6. Also, examples of statistical tests that provide either sufficient or insufficient information to calculate effect sizes and statistical significance ourselves are included in Appendix-6. You can usually find information for Item 5 in the Results.

(6) *Result(s) of statistical analysis/analyses for outcome measure.* Please code the result of the statistical analysis or analyses on the outcome measure you are coding. Results fall into any of the following categories:

> **(1)** significant difference, with treatment outperforming controls (the treatment examined has scores that are statistically significantly better than the control condition)
> **(2)** significant difference, with treatment under-performing controls (the treatment examined has scores that are statistically significantly worse than the control condition)
> **(3)** no statistically significant difference between conditions

You can usually find information for Item 6 in the Results section. Very important, in cases in which statistical significance and effect size need to be estimated for an outcome finding, please consult Appendix-7 for methods that you will use to calculate statistical tests and effect sizes. **Note. Item 6 deals with what the final outcome was for the outcome measure. This item deals specifically with what the main finding was (i.e., were the conditions different, if so, how).**

VERY IMPORTANT: FOR ITEM 6, YOU WILL HAVE TO CALCULATE STATISTICAL SIGNIFICANCE YOURSELF. PLEASE CONSULT APPENDIX-7 FOR ONLINE CALCULATORS TO CALCULATE STATISTICS IN THESE INSTANCES.

(7) *Statistical information of results of statistical analysis for outcome measure.* For each outcome measure, please code the result of the statistical analysis. The result usually consists of the score of the test (e.g., chi square[1, N = 96] = 2.81, t[48] = 3.67), along with whether the *p*-value was below a given cutoff (e.g., *p* < .05). In any case, please code all the statistical information resulting from your statistical calculations using the online calculators. **When you estimate statistical significance using data provided in the article, please provide this info (please consult Appendix-7). Sometimes you may be calculating an effect size based on test statistics, like *t* test, because the only information available to calculate an effect size is a test statistic. In those cases, please code this statistical test data (including *p*-value) for Item 7.**

Note. The following Items (8-15) should be coded across all outcome measures, methods of examination (e.g., mean differences, diagnostic status) and types of statistical tests (e.g., *t* tests, chi square). All other items in this section should be coded within each individual finding used in the study to examine outcomes in the construct directly targeted for treatment. By each individual finding, what we mean is each individual outcome finding, or each instance in which the authors of the article compares the treatment condition to the control condition. That is, the same outcome <u>measure</u> can be examined in multiple ways in a given article (yielding multiple findings), such as: (a) a questionnaire being examined more than once because it has multiple subscales; or (b) a measure being examined for mean differences between treatment and control conditions, as well as for differences between conditions in clinically significant change. Thus, in this section, each outcome finding means each time an outcome measure in the article was examined.

(8) *Total number of outcome measures administered of the construct targeted for treatment.* Please code how many measures of the construct targeted for treatment for which the authors had data available for them post-treatment. This data can be available in a variety of formats (e.g., means and standard deviations, frequencies, results from a statistical test [*F test*, *t test*, chi square]. If there is ANY data available for a given measure of the problem targeted for treatment, for which there is sufficient information to calculate effect sizes (please see Appendix-6), please count that outcome measure in the total number. The list of available outcome measures for CBT for child anxiety and BPT for child conduct are available in Appendices 4 and 5, respectively. You can usually find this information using a combination of the Method (usually in a Measures subsection of the Method), and Results (usually by looking over outcome data provided by the author in the Results). **Please be careful to note that when the same outcome measure is administered to more than one informant, each informant's measure should count as a separate, single outcome measure.**

Note. Item 8 has to do with outcome measures and not outcome <u>findings</u>. For example, a measure may be used more than one time to examine treatment outcomes (e.g., on structured interview, diagnostic status, and then also clinician severity ratings taken from diagnosis taken from structured interview); these are different

findings, but the same measure. Further, a measure may have multiple subscales reported for it, and each subscale may be examined at post-treatment to gauge treatment outcomes. However, each subscale does not count as an outcome measure, only an outcome finding (more on outcome findings below). Please only count the entire measure from which the subscales are derived as a single outcome measure. Thus, in Item 8, please code for number of different measures of outcome used in the study to assess outcomes in the construct targeted for treatment.

(9) *Total number of methods of analysis that were employed to examine outcome measures of the construct targeted for treatment*. Please code how many methods of analysis were employed to examine outcome measures of the construct targeted for treatment. This code should be applied across outcome measures. In other words, methods may not be applied to all outcome measures. For instance, tests of diagnostic status may only be applied to one particular measure (e.g., diagnoses derived from a structured interview), whereas tests of mean differences may be employed to examine almost all of the measures. The tests that you may see usually fall into three categories: **(a)** tests of mean differences; **(b)** tests of diagnostic status; and **(c)** tests of clinically significant change. A helpful note with this is that you can usually find out whether tests of mean differences were used by looking at the Tables in the article. However, tests of diagnostic status and clinically significant change are usually reported in the text of the article. Examples of the different methods of analysis are illustrated in Appendix-6. You can usually find this information in the Results. **Note. Some methods (e.g., clinically significant change) combine multiple outcome measures to gauge change. First, please make sure that each method used in the study that is included in the meta-analysis only includes information gathered immediately post-treatment (i.e., it does not include information gathered at some follow-up period after post-treatment assessment). Second, please make sure that each method used in the study that is included in the meta-analysis only includes information gathered from measures listed in Appendices 4 and 5.**

(10) *Total number of types of statistical tests that were employed to examine outcome measures of the construct targeted for treatment*. Please code how many types of statistical tests were employed to examine outcome measures of the construct targeted for treatment. This code should be applied across outcome measures, and should be based on the statistical tests that you coded above (Item 5, Appendix-7). In other words, types of statistical tests may not be applied to all outcome measures. For instance, when treatment and control conditions are compared on diagnostic status, this test may only be examined using chi square, whereas tests of mean differences between conditions may be examined using *t* tests. The tests that you may use fall into two categories: **(a)** *t* tests; and **(b)** chi square. A helpful note with this is that you can usually find out whether you should use *t* tests for an outcome finding by looking at the Tables in the article. However, data you may use to calculate chi squares are usually reported in the text of the article. Examples of the different types of statistical analysis are illustrated in Appendix-6. You can usually find this information in the Results.

(10a) *Completer Analysis or Intent-to-Treat?* Across all outcome measures, methods of examining outcome measures, and methods of statistical analysis employed in the study to examine outcomes in the construct targeted for treatment, please code whether treatment outcomes were examined using Completer Analyses (i.e., only the people providing outcome data at the end of treatment were examined; please code this **1**) or whether the authors employed Intent-to-Treat analyses to examine outcome (i.e., all participants were used to examine outcome, regardless of dropout, with the most recent data point used [usually pre-treatment scores if dropout occurs over course of treatment]; please code this **2**). In cases in which both completer and intent-to-treat analyses were reported in the study, please code the study in terms of which analyses were reported in the study with the most data (i.e., which analyses do the authors include means and standard deviations, frequencies, and not just results of statistical analyses). **Very Important: Please make sure that the types of results coded in Items 1-7 match the code made in Item 10a (i.e., completer findings coded if Item 10a is coded "1").**

(11) *Total number of information sources relied on to measure outcomes in the construct targeted for treatment*. Across all outcome measures employed in the study to examine outcomes in the construct targeted for treatment, please code how many informants were employed to complete these outcome measures. This code should be applied across outcome measures, and this information can usually be found with a combination of the Methods and Results sections of the article. Also, much of the information needed to code this item has been entered or used in previous codes, so you could use the work you have done already to gather this information. Again, informants will normally fall into the following categories:

Information Source	**What Informant Type Counts?**
Child self-report	Child
Parent report, **Mother**; Parent report, **Father**	Parent **(Although mother and father each separately count as a single informant)**
Teacher report	Teacher
Independent observer	Independent observer
Clinical interviewer based on child report	Interviewer AND Child
Clinical interviewer based on parent report	Interviewer AND Parent
Clinical interviewer based on teacher report	Interviewer AND Teacher
Clinical interviewer based on multiple reports (e.g., parent/child; parent/teacher)	Interviewer AND ANY informant relied on by interviewer

In addition, as before please keep in mind that in a given study, multiple outcome measures may be coded for the multiple information sources above, because the same structured interview may be administered to multiple people in a study. However, not all of these measures may have information about them in the Results, as far as whether they were actually used after the intervention to assess outcomes. For example, in a child anxiety intervention study, a structured interview may have been given by a clinician to both parent and child before treatment, but outcomes on that structured interview may only be provided after treatment based on clinicians' interviews of parents. Thus, as in all items in this section, please limit including informants in the total number of informants to those informants that were employed in the study to gather information or assess outcomes after the intervention was complete.

Note. For Item 11 (and all other items for which information sources are of interest), in the case of clinician interviews based on information provided by a specific source or sources, please count that interviewer measure as both a clinical interviewer measure AND a measure of the information source upon which the interview is based. For example, a clinical interview based on the parent as the source counts as a clinical interviewer AND a parent measure for ratings.

Another Note. Sometimes the same measure is reported on by multiple informants of the same category, like both mother and father completing the same scale. Both mother and father count as "parent," in the informant type category, but each should count as a different informant number in the totals done in Item 11. For instance, if both mother and father are completing the same scale, the mother counts as an informant in the total number of informants, and the father counts as another informant in the total number of informants.

(12) *Total number of outcome measure methodologies employed to measure outcomes in the construct targeted for treatment*. Across all outcome measures employed in the study to examine outcomes in the construct targeted for treatment, please code how many different methods of measurement were employed to assess outcomes. This code should be applied across outcome measures, and this information can usually be found with a combination of the Methods and Results of the article. Also, much of the information needed to code this item has been entered or used in previous codes, so you could use the work you have done already to gather this information. Again, outcome measure methodologies will normally fall into the following categories: **(a)** questionnaire; **(b)** behavioral observation made in a lab or home (usually by an independent observer); and **(c)** structured interview.

(13) *Total number of findings examining outcomes in the construct targeted for treatment that were statistically significant.* Across all outcome measures, methods of examining outcome measures, and methods of statistical analysis employed in the study to examine outcomes in the construct targeted for treatment, please code how many total findings gleaned from these examinations suggested that the treatment and control condition(s) were statistically significantly different from one another. Please only include in this coding the number of findings for which there is enough information to both calculate whether treatment conditions were statistically significantly different from one another, as well as calculate effect sizes. Also, please include findings in this total that suggest **ANY** statistical difference, regardless of whether that difference suggests that treatment was better than controls, or control was better than treatment. You can usually find information for this code with a combination of information provided in text and tables in the Results. As before, please note that a given outcome measure can be examined in multiple ways, and each of these multiple ways of examining an outcome measure counts as a finding (provided, of course that enough information was provided in the article to infer statistical differences between conditions, and calculate effect sizes). The coding sheet provides a table for you to write in the necessary information to gauge how many findings in the study relate to examining post-treatment changes in the construct primarily targeted for treatment, and how many were statistically significant. As mentioned previously, there will be instances in which an outcome measure that assesses changes in the construct targeted for treatment was seemingly employed to examine post-treatment changes, but insufficient information is provided by the author to calculate effect sizes and statistical significance ourselves. Thus, before you include a given finding within the total number of findings employed within the study, please check carefully whether this finding was reported in the article in a way for which: **(a)** statistical results of the finding can be calculated; and **(b)** effect sizes can be calculated. Please consult Appendix-7 for methods that you will use to calculate effect sizes and infer statistical significance. Also, please consult Appendix-6 for examples of results reported in articles where there is sufficient and insufficient information to calculate effect sizes and statistical significance.

(14) *Total number of findings examining outcomes in the construct targeted for treatment.* Across all outcome measures, methods of examining outcome measures, and methods of statistical analysis employed in the study to examine outcomes in the construct targeted for treatment, please code how many total findings were gleaned from these examinations, regardless of whether the finding was statistically significant or not. Please only include in this coding the number of findings for which there is enough information to both calculate directly from the article whether treatment conditions were statistically significantly different from one another, as well as calculate effect sizes. As mentioned previously, there will be instances in which an outcome measure that assesses changes in the construct targeted for treatment was seemingly employed to examine post-treatment changes, but insufficient information is provided by the author to calculate effect sizes and statistical significance. Thus, before you include a given finding within the total number of findings employed within the study, please check carefully whether this finding was reported in the article in a way for which: **(a)** statistical results of the finding can be calculated; and **(b)** effect sizes can be calculated. Information for this coding can usually be found with a combination of the Methods and Results sections of the article. Also, much of the information needed to code this item has been entered or used in previous codes, so you could use the work you have done already to gather this information. As before, please note that a given outcome measure can be examined in multiple ways, and each of these multiple ways of examining an outcome measure counts as a finding (provided, of course that enough information was provided in the article to calculate statistical differences between conditions, and calculate effect sizes). The coding sheet provides a table for you to write in the necessary information to gauge how many findings in the study relate to examining post-treatment changes in the construct primarily targeted for treatment. **Note. A single measure can contribute more than one finding if: (1) the measure is examined using more than one method; and/or (2) more than one subscale of the measure is examined using one or more methods.**

(15) *Percentage of findings examining outcomes in the construct targeted for treatment that were statistically significant (Please show work [i.e., number significant findings divided by total number of findings]).* Across all outcome measures, methods of examining outcome measures, and methods of statistical analysis employed in the study to examine outcomes in the construct targeted for treatment, please code the percentage of findings examining outcomes in the construct targeted for treatment that were statistically significant. For this coding, please show all mathematical work used to find the percentage (basically Item 13 divided by Item 14).

VII. Data Collection Part 3: Coding Information to Calculate Effect Sizes

In this section, we are interested in coding information that will be required to calculate effect sizes for the findings we coded in the previous section. We are going to code both the raw data information that is utilized to calculate effect sizes, as well as calculate the effect sizes themselves. We will calculate effect sizes one way for each finding now, and code the raw data in case we need to calculate effect sizes another way later on. Because we are going to be entering all this information in a software program that calculates effect sizes for us, if we need to recalculate effect sizes from the raw data, software can do that for us later on. Details are provided below on what information to put down for each kind of method of statistical analysis (see Appendix-7).

Note. Before you begin this section, it is important for you to know that not all studies will conduct tests for all of the kinds of treatment comparisons you see below (i.e., mean differences, diagnostic status, clinically significant change). For example, a study may conduct tests of mean differences between treatment and control conditions, but not for diagnostic status or clinically significant change. Also, it is just as important to know that outcome measures may be examined in more than one way. For example, sometimes the means of both groups of ratings gathered from the total score of a questionnaire measure may be compared using a *t* test of comparisons of means, and that same total score from that measure may also be examined using tests of clinically significant change (e.g., comparisons of groups in terms of the numbers of children that have total scores that fall below a given cutoff). In sum, please keep in mind that the same score from the same measure can be examined in various ways, so please search the article carefully to make sure that you catch all of the comparisons between conditions of interest. Most importantly, please calculate effect sizes relying primarily on group data (e.g., means and standard deviations, numerical frequencies), and only use other forms of data (e.g., statistical test data such as *t*-values, *p*-values) if the data analyzed to arrive at the statistical results are not provided. Please refer to Appendix-7 for the formulas or calculations to perform for any of the effect sizes described in this section.

(1a) *Coding pre-treatment equivalence using statistical tests and effect size information on pre-treatment scores for outcomes.* For each outcome finding, please code statistical test and effect size information for pre-treatment score comparisons between the treatment and control conditions for which effect sizes are being calculated. On the basis of this information, please also code "yes" (equivalent) or "no" (not equivalent) with regard to whether the treatment and control conditions being compared were equivalent prior to treatment on the outcome finding being coded. **The pre-treatment equivalency codes are made in three steps. First,** for the statistical tests on pre-treatment equivalence, please consult Appendix-7. Remember to please conduct the statistical test for pre-treatment equivalency that corresponds to the effect size or effect sizes you are conducting for the outcome (e.g., if means and standard deviations: *t* test; if frequencies: chi square, if both on same outcome, then calculate both for that outcome). Please report the test statistic as well as the *p*-value. **Second,** please calculate a pre-treatment effect size for each outcome for which a post-treatment effect size will be calculated. Specifically, please calculate an **unadjusted** Glass's Δ effect size for mean differences outcomes coded using Glass's Δ, and please calculate an **unadjusted** Cohen's *d* effect size for outcomes for which the final outcome being coded utilizes the Cohen's *d* metric (e.g., any outcome coded using test statistics like *t*, *F*, or chi square; diagnostic status and clinically significant change findings calculated using diagnostic or change frequencies of participants). Please refer to Appendix-7 for methods of calculating unadjusted Glass's Δ, unadjusted Cohen's *d* using test statistics, and unadjusted Cohen's *d* using diagnostic status and clinically significant change findings. **Third,** please answer "yes" (**1**) or "no" (**2**) as to whether the treatment and control condition were equivalent pre-treatment on the outcome or outcomes for which post-treatment effect sizes will be calculated. Specifically, please code "yes" (groups equivalent prior to treatment) if the pre-treatment statistical test was not significant **AND** the pre-treatment effect size (Glass's Δ, Cohen's *d*) was below .20. Please code "no" (groups not equivalent prior to treatment) if the pre-treatment statistical test was significant **OR** the pre-treatment effect size (Glass's Δ, Cohen's *d*) was greater than or equal to .20.

(1) *Coding effect size information for findings based on tests of mean differences between conditions.* Effect size findings based on tests of mean differences between conditions will generally be coded one way: Using a measure referred to as Glass's Δ. Specifically, for any finding based on comparisons of mean differences in which the author provides both the means and standard deviations of both experimental conditions, you would just take the post-treatment control group mean and subtract it from the post-treatment intervention group mean (the intervention being examined in the study), and then divide the resulting number from the *control group's* post-treatment standard deviation (see Appendix-7 for formula and directions). Please code an effect size, as well as all raw data information (means and standard deviations for **BOTH** experimental conditions being compared) for each finding you coded information on in the previous section. Although only the control group's standard deviation will be used to calculate the effect size now, other effect sizes that we may calculate later on use both experimental conditions' standard deviations when determining effect sizes. Please refer to your work in the previous section to determine for which findings effect sizes ought to be calculated. Please also note again that a given treatment study may be interested in comparing more than one treatment to a control group, or may even compare a treatment group or two to more than one control condition. Thus, for each effect size, it is important that you note on the coding sheet which groups were being compared to derive the effect size. Thankfully, codes you made in the previous section can assist you in this, and the coding sheet is constructed to aid you in clearly providing this information. Please contact me if you have any questions on this or other items in this section.

(2) *Coding effect size information for findings based on tests of diagnostic status between conditions.* Coding effect size information for findings based on tests of diagnostic status will be calculated using fundamentally different effect size metrics, relative to the previous code. Specifically, these tests compare the frequencies of participants in each group that either do or not meet diagnostic criteria for the problem primarily targeted for treatment. Thus, this code relies on frequencies of participants that fall into one group or another on a dichotomous variable. For these effect sizes, we will use an effect size measure that is called the Phi (Φ) coefficient (see Appendix-7 for formula and directions). First, to calculate this measure, we need the article to provide the raw frequencies of participants, and not just percentages, because sometimes treatment attrition and missing data make it difficult to estimate frequencies from percentages. Also, please keep in mind that for these tests, sometimes treatment studies will test diagnostic status by pooling all of the treated participants into one group, and comparing this group to the control condition. Regardless of the composition of this group (e.g., both of the treatment groups are treatments for which we are interested in gathering effect sizes), we cannot use information derived from this test to calculate effect sizes. The frequencies of participants in each treatment and control condition in the study must be reported separately in the article for us to calculate effect sizes for these findings. Also, sometimes frequencies are not provided in the article, and the only information provided about findings of clinically significant change is statistical information (e.g., *t*, chi square). In these instances, please note that this information was used to calculate the effect size, and please refer to Appendix-7 for formulas to calculate effect sizes using test statistics. Please refer to your work in the previous section to determine for which findings effect sizes ought to be calculated. Please also note again that a given treatment study may be interested in comparing more than one treatment to a control group, or may even compare a treatment group or two to more than one control condition. Thus, for each effect size, it is important that you note on the coding sheet which groups were being compared to derive the effect size.

(3) *Coding effect size information for findings based on tests of clinically significant change between conditions.* Coding effect size information for findings based on tests of clinically significant change will be calculated using similar effect size metrics, relative to tests of diagnostic status. Specifically, these tests compare the frequencies of participants in each group that either do or not pass thresholds for achieving clinically significant change in the problem primarily targeted for treatment. Thus, like diagnostic status, this code relies on frequencies of participants that fall into one group or another on a dichotomous variable. For these effect sizes, we will use Φ (see Appendix-7 for formula and directions). First, to calculate this measure, we need the article to provide the raw frequencies of participants, and not just percentages, because sometimes treatment attrition makes it difficult to estimate frequencies from percentages. Also, please keep in mind that for these tests, sometimes treatment studies will test clinically significant change by pooling all of the treated participants into one group, and comparing this group to the control condition. Regardless of the composition of this group (e.g., both of the treatment groups are treatments for which we are interested in gathering effect sizes), we cannot use information derived from this test to calculate effect sizes. The frequencies of participants in each treatment and control condition in the study must be reported separately in the article for us to calculate effect sizes for these findings. Also, sometimes frequencies are not provided in the article, and the only information provided about findings of clinically significant change is statistical information (e.g., *t*, chi square). In these instances, please note that this information was used to calculate the effect size, and please refer to Appendix-7 for formulas to calculate effect sizes using test statistics. Please refer to your work in the previous section to determine for which findings effect sizes ought to be calculated. Please also note again that a given treatment study may be interested in comparing more than one treatment to a control group, or may even compare a treatment group or two to more than one control condition. Thus, for each effect size, it is important that you note on the coding sheet which groups were being compared to derive the effect size.

Note. Very important, tests of diagnostic status are their own method of analysis (for ANY of these tests, please consult Item 2 of this section above), and do not count as a test of clinically significant change, even if the authors identify tests of diagnostic status as a test of clinically significant change.

(4) *Groups equivalent prior to treatment on all outcome measures of construct targeted for treatment?* Across all outcome measures and methods of examining outcomes, please code whether the treatment and control groups being compared were equivalent prior to treatment on **ALL** of the outcome measures coded above (Equivalent = **1**; Not Equivalent = **2**). Again, you will know whether groups are equivalent prior to treatment from two sources. First, you may look back on your previous calculations of tests of significance on all of the pre-test scores of the outcome findings coded previously (please consult online calculator formulas in Appendix-7). If the groups being compared on these findings are significantly different from each other prior to treatment on **ANY** of the significance tests you conducted, please code No (**2**).

The second and more difficult source of examining equivalence is to calculate pre-treatment effect sizes for the groups being compared; please consult Appendix-7 and this section of the manual (Part 3) to determine what methods to employ to calculate pre-treatment effect sizes. You can usually find pre-treatment data at the very beginning of the Results, and/or in a table of pre- and post-treatment data for the groups being compared. Please double-check to make sure effect sizes are calculated on pre-treatment values; sometimes it is easy to confuse pre-treatment scores with post-treatment scores because they are often in the same table. By Item 4, you should have already calculated all pre-treatment statistical tests and effect sizes for the outcome measures coded in this section, so please check back to see if these effect sizes are at least or above unadjusted Glass's Δ or unadjusted Cohen's *d* = .20 (please consult Appendix-7, because some effect sizes have to be calculated first using an "*r*" metric effect size, and then converted to "*d*"). If ANY of the effect sizes is at least or exceeds Glass's Δ or *d* = .20, please code Item 4 as No (**2**).

Note on deciding on whether groups are equivalent. To be clear, for each outcome measure coded in this section, please examine **BOTH** statistical significant tests and effect size indices to see if groups are equivalent. If the groups are non-equivalent on **EITHER** statistical significance or effect size measures on **ANY** of the outcome measures, then Item 4 is coded **NO** (**2**).

NOTE ON EXAMINING PRE-TREATMENT EQUIVALENCE. PLEASE NOTE THAT EQUIVALENCE BETWEEN OUTCOME MEASURES DEPENDS ON OUTCOME MEASURE RATINGS PRIOR TO TREATMENT, BUT ALSO ON THE METHOD FOR WHICH ONE IS EXAMINING PRE-TREATMENT EQUIVALENCE (E.G., MEAN DIFFERENCES BETWEEN CONDITIONS; DIAGNOSTIC STATUS). FOR INSTANCE, A DIAGNOSTIC INTERVIEW MEASURE OF ANXIETY MAY BE STATISTICALLY EQUIVALENT BETWEEN CONDITIONS PRIOR TO TREATMENT IF THE GROUP MEANS ON SYMPTOM RATINGS ARE COMPARED, BUT NOT IF DIAGNOSTIC STATUS IS COMPARED. THUS, FOR EACH OUTCOME MEASURE BEING EXAMINED FOR EQUIVALENCE, PLEASE EXAMINE EQUIVALENCE ON THE MEASURE USING EACH METHOD BEING USED IN THE STUDY TO EXAMINE THAT OUTCOME MEASURE. FOR INSTANCE, IF THE OUTCOME MEASURE BEING EXAMINED IN THE STUDY IS EXAMINED USING BOTH MEAN DIFFERENCES AND ANOTHER METHOD OF EXAMINING OUTCOME, PLEASE EXAMINE PRE-TREATMENT EQUIVALENCE ON THE OUTCOME MEASURE ON BOTH METHODS.

(5) *If non-equivalence observed (Item 4 = 2), how many outcome measures were non-equivalent?* If groups are non-equivalent on any of the outcome findings of the construct targeted for treatment, please code how many of the outcome findings were non-equivalent prior to treatment.

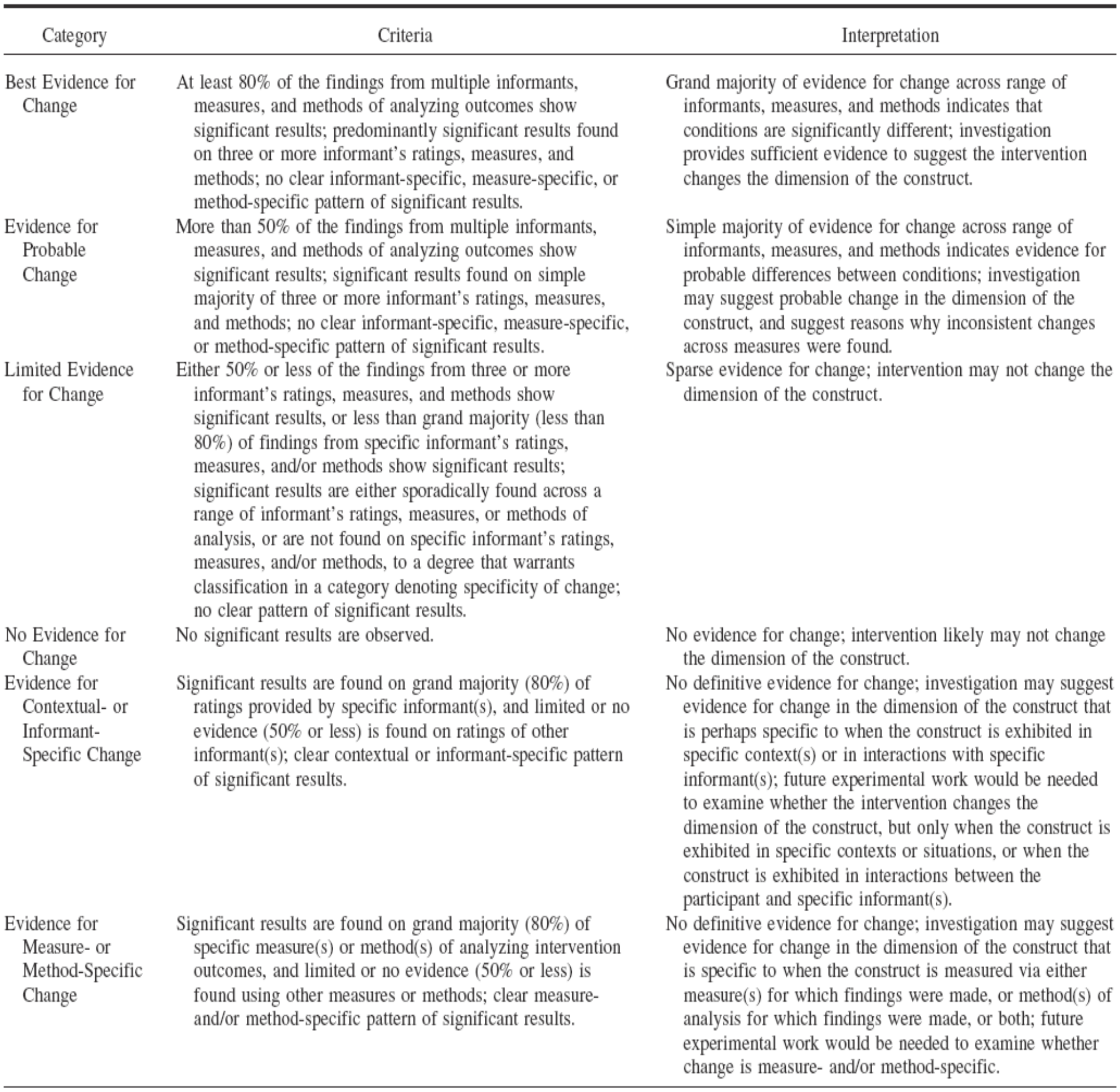
VIII. Data Collection Part 4: Coding Classifications of Studies

In this section, we will be taking information coded previously in each article to classify studies in terms of the evidence they provide for the treatments they examine, and coding information on manualization of treatments. We will be coding classifications of studies using two systems. Each system will be described in detail below, and instructions will be provided on how to employ each system to arrive at classifications of studies for the treatment or treatments they examine. Some codes made in Part 4 take into account the totality of information coded across all studies included in the meta-analysis. Thus, Part 4 should be coded only after all of the information from Parts 1 through 3 is coded for all studies.

Note. As in previous sections, please remember that a given article may be classified more than once using either system. For example, an article may compare a single treatment to more than one control condition. In this instance, the treatment may be classified more than once, although this may only apply to one of the classification systems (see below). Another example may be a study that examines two treatments separately that meet criteria for inclusion in the study (see Data Collection Part 1, Code 5; Appendices 2 and 3), and compares each of these treatments to a control condition. In this second example, the evidence provided when examining each treatment separately should each be classified separately in its own proper code.

Classification of studies under System 1. Coding for classifications of studies under System 1 will be based on the system noted in the table below. Classifications in this system rely on the percentage of comparisons between a specific treatment compared to a specific control condition that are significantly different, the characteristics of these significant differences, and the effect sizes of these differences. What can be evident from this system is that this system uses all of the findings gathered from the study that were used to examine whether the intervention was significantly different from the control condition on measures of the problem primarily targeted for treatment. As mentioned previously, information used to classify studies based on this system has already been collected in prior sections. This coding compiles this information to arrive at a single classification coding for a specific treatment-control comparison. Thus, within a given study, multiple classification codes may be made, depending on how many treatment conditions that met criteria for inclusion in the meta-analysis were examined in the article, and how many control groups were included in the meta-analysis. Additionally, multiple classification codes can be made when a treatment is compared to a control condition, and the findings from this comparison meet criteria for more than one classification coding (please see note below on multiple classifications). A table outlining the different classification codes that may be made under System 1 is provided below, as well as a more detailed description of each System 1 classification code.

Note. Under System 1, a critical factor involves gauging whether a range of different information sources (at least three or more) were employed in a study, or whether a pattern of significant findings in a study includes a minimum number of a particular information source (e.g., at least three parent-rated outcomes were used in the study, and all three measures suggest the treatment and control condition were significantly different). However, often times an outcome measure is employed that relies on two different sources to gather information on that measure. For instance, information gathered from a structured interview outcome measure is usually collected by a clinical interviewer, based on the information provided from a particular source or sources (e.g., interview based on parent report, interview based on both parent and child report). Under these circumstances, information gleaned from the measure cannot be gleaned without one of the two sources. Stated another way, each source essentially is equally responsible for the information gleaned from that measure. In these cases, please count outcome measures where one informant is providing information on that measure, based on the report of another informant as an outcome measure completed by BOTH informants. Thus, in System 1 classifications, outcome measures using more than one informant count as one measure completed by each informant involved in gathering information on that measure. For example, a measure of diagnostic status taken from a structured interview completed by a clinical interview, and based on parent report counts as a clinical interviewer measure AND a parent report measure.



Category	Criteria	Interpretation
Best Evidence for Change	At least 80% of the findings from multiple informants, measures, and methods of analyzing outcomes show significant results; predominantly significant results found on three or more informant's ratings, measures, and methods; no clear informant-specific, measure-specific, or method-specific pattern of significant results.	Grand majority of evidence for change across range of informants, measures, and methods indicates that conditions are significantly different; investigation provides sufficient evidence to suggest the intervention changes the dimension of the construct.
Evidence for Probable Change	More than 50% of the findings from multiple informants, measures, and methods of analyzing outcomes show significant results; significant results found on simple majority of three or more informant's ratings, measures, and methods; no clear informant-specific, measure-specific, or method-specific pattern of significant results.	Simple majority of evidence for change across range of informants, measures, and methods indicates evidence for probable differences between conditions; investigation may suggest probable change in the dimension of the construct, and suggest reasons why inconsistent changes across measures were found.
Limited Evidence for Change	Either 50% or less of the findings from three or more informant's ratings, measures, and methods show significant results, or less than grand majority (less than 80%) of findings from specific informant's ratings, measures, and/or methods show significant results; significant results are either sporadically found across a range of informant's ratings, measures, or methods of analysis, or are not found on specific informant's ratings, measures, and/or methods, to a degree that warrants classification in a category denoting specificity of change; no clear pattern of significant results.	Sparse evidence for change; intervention may not change the dimension of the construct.
No Evidence for Change	No significant results are observed.	No evidence for change; intervention likely may not change the dimension of the construct.
Evidence for Contextual- or Informant-Specific Change	Significant results are found on grand majority (80%) of ratings provided by specific informant(s), and limited or no evidence (50% or less) is found on ratings of other informant(s); clear contextual or informant-specific pattern of significant results.	No definitive evidence for change; investigation may suggest evidence for change in the dimension of the construct that is perhaps specific to when the construct is exhibited in specific context(s) or in interactions with specific informant(s); future experimental work would be needed to examine whether the intervention changes the dimension of the construct, but only when the construct is exhibited in specific contexts or situations, or when the construct is exhibited in interactions between the participant and specific informant(s).
Evidence for Measure- or Method-Specific Change	Significant results are found on grand majority (80%) of specific measure(s) or method(s) of analyzing intervention outcomes, and limited or no evidence (50% or less) is found using other measures or methods; clear measure- and/or method-specific pattern of significant results.	No definitive evidence for change; investigation may suggest evidence for change in the dimension of the construct that is specific to when the construct is measured via either measure(s) for which findings were made, or method(s) of analysis for which findings were made, or both; future experimental work would be needed to examine whether change is measure- and/or method-specific.

Specific criteria for System 1, Best Evidence for Change. This first classification category is reserved for those study comparisons between treatment and control that find significant differences on at least 80% of all findings, and these findings were based on: **(a)** three outcome measures each from at least three different informants (e.g., parent, child, teacher); **(b)** measurements taken using three methodologically distinct outcome measure types (e.g., questionnaire, structured interview, laboratory observations); and **(c)** examinations of outcomes using three distinct methods of analysis (e.g., mean differences between conditions, diagnostic status, clinically significant change). In other words, this category is for studies that employ ranges of outcome findings that each taps into a range of ways of gauging outcomes, and this totality of outcome findings reveals significant differences between conditions at least 80% of the time. Note that this category, like all categories, focuses on *differences* between conditions, and not necessarily findings that suggest the treatment condition "outperformed" the control condition. Significant differences between conditions could reflect that the control condition had better outcomes, relative to the treatment condition. You can think of this category in terms of a checklist of outcome findings, in which you can check off consistency of significant results for a particular way of gauging outcome on three of each of the three major ways in which outcome findings may differ: **(a)** the informant relied on to measure the outcome; **(b)** the methodology employed to measure the outcome; and **(c)** the statistical method employed to examine the outcome. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X		X
Child	Questionnaire	X		
Child	Structured Interview	X	X	
Parent	Questionnaire	X		X
Parent	Questionnaire	X		X
Parent	Questionnaire	–		X
Parent	Structured Interview	–	X	
Parent	Structured Interview	X	X	
Lab Observer	Independent Observation			X
Lab Observer	Independent Observation	X		X
Lab Observer	Independent Observation	X		

Where "X" denotes a significant difference between treatment and control conditions, "–" denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. In this example, there were a couple of non-significant differences in the table, but these non-significant differences were not enough to pull a particular way of gathering outcomes under the 80% threshold of significant differences. For instance, all of the child-rated outcomes, diagnostic status findings, and clinically significant change findings are significant. These findings aid in meeting criteria for this category. Further, some of the parent-rated, mean differences, questionnaire, and structured interview findings are non-significant, but no single way of gathering outcome findings falls under the 80% threshold. Indeed, mean differences findings are 80% significant, parent-rated findings are 80% significant, questionnaire findings are 89% significant, and structured interview findings are 83% significant. Finally, as can be readily seen, this category is reserved for studies that compare treatment and control conditions with a wide range of ways of gathering and examining outcome findings, and this wide range of outcome findings yields predominantly significant differences between treatment and control conditions.

Specific criteria for System 1, Evidence for Probable Change. This second classification coding is reserved for those study comparisons between treatment and control that find significant differences on more than 50% (i.e., simple majority) of all findings, and these findings were based on: **(a)** three outcome measures each from at least three different informants (e.g., parent, child, teacher); **(b)** measurements taken using three methodologically distinct outcome measure types (e.g., questionnaire, structured interview, laboratory observations); and **(c)** examinations of outcomes using three distinct methods of analysis (e.g., mean differences between conditions, diagnostic status, clinically significant change). Like the first classification category, there should be no informant, measure, or method-specific pattern of significant findings. This category is for studies that employ ranges of outcome findings that each taps into a range of ways of gauging outcomes, and this totality of outcome findings reveals significant differences between conditions at least above 50% of the time. Note that this category, like all categories, focuses on *differences* between conditions, and not necessarily findings that suggest the treatment condition "outperformed" the control condition. Significant differences between conditions could reflect that the control condition had better outcomes, relative to the treatment condition. You can think of this category in terms of a checklist of outcome findings, in which you can check off a simple majority of significant results for a particular way of gauging outcome on three of each of the three major ways in which outcome findings may differ: **(a)** the informant relied on to measure the outcome; **(b)** the methodology employed to measure the outcome; and **(c)** the statistical method employed to examine the outcome. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X		X
Child	Questionnaire	–		
Child	Structured Interview	–	X	
Parent	Questionnaire	X		X
Parent	Questionnaire	X		–
Parent	Questionnaire	–		X
Parent	Structured Interview	–	X	
Parent	Structured Interview	X	X	
Lab Observer	Independent Observation			–
Lab Observer	Independent Observation	X		X
Lab Observer	Independent Observation	X		

Where "X" denotes a significant difference between treatment and control conditions, "–" denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. In this example, there were a few more non-significant differences in the table, relative to the "Best Evidence for Change" category, but these non-significant differences were not enough to pull a particular way of gathering outcomes under the simple majority (above 50%) threshold of significant differences. Indeed, each of the different ways of gauging treatment outcomes is above 50% on significant outcomes. For instance, mean differences findings are 60% significant, parent-rated findings are 70% significant, questionnaire findings are 67% significant, and structured interview findings are also 67% significant. The rest of the outcome finding permutations (child, lab observer, independent observation, diagnostic status, clinically significant change) is at least above the 50% threshold. Keep in mind that a particular way of gauging outcomes may be well above 50% in this classification category. For instance, diagnostic status findings in the table are all significant. However, all of the other ways of gauging outcomes are at least above 50%, making this study properly classified in this System 1 category. Finally, as can be readily seen, this category is reserved for studies that compare treatment and control conditions with a wide range of ways of gathering and examining outcome findings, and this wide range of outcome findings yields a simple majority of significant differences between treatment and control conditions. Again, no specificity of predominantly significant results ought to be observed.

Specific criteria for System 1, Limited Evidence for Change. The "Limited Evidence for Change" classification is reserved for those study comparisons between treatment and control that find significant differences on either: **(1)** less than 50% of all findings, and these findings were based on: **(a)** three outcome measures each from at least three different informants (e.g., parent, child, teacher); **(b)** measurements taken using three methodologically distinct outcome measure types (e.g., questionnaire, structured interview, laboratory observations); and **(c)** examinations of outcomes using three distinct methods of analysis (e.g., mean differences between conditions, diagnostic status, clinically significant change); or **(2)** less than a grand majority (less than 80%) of findings gathered from specific informant's ratings, measures, and/or methods of analysis. Similar to the first two classification categories, a study could be classified in this category when there is no informant, measure, or method-specific pattern of significant findings. However, a study can be classified in this category if limited information sources, measures, and/or methods were employed in a study (i.e., less ranges of ways in which outcomes can be examined), and this limited range of outcome examinations yields a pattern or patterns of findings that do not suggest informant-, measure-, and/or method specificity in significant findings (see "Evidence for Informant-Specific Change" and "Evidence for Measure- or Method-Specific Change" category criteria). Note that this category, like all categories, focuses on *differences* between conditions, and not necessarily findings that suggest the treatment condition "outperformed" the control condition. Significant differences between conditions could reflect that the control condition had better outcomes, relative to the treatment condition. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	–		–
Child	Questionnaire	–		
Child	Structured Interview	X	X	
Parent	Questionnaire	X		–
Parent	Questionnaire	X		–
Parent	Questionnaire	–		X
Parent	Structured Interview	–	–	
Parent	Structured Interview	X	–	
Lab Observer	Independent Observation			–
Lab Observer	Independent Observation	–		X
Lab Observer	Independent Observation	–		

Where “X” denotes a significant difference between treatment and control conditions, “–“ denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. In this example, there were far more non-significant differences in the table, relative to the “Best Evidence for Change” and “Probable Evidence for Change” categories, and these non-significant differences are enough to pull every particular way of gathering outcomes **UNDER** the simple majority (above 50%) threshold of significant differences. Indeed, each of the different ways of gauging treatment outcomes is less than or equal to 50% on significant outcomes. For instance, mean differences findings are 40% significant, parent-rated findings are 40% significant, questionnaire findings are 33% significant, and structured interview findings are 33% significant. The rest of the outcome finding permutations (child, lab observer, independent observation, diagnostic status, clinically significant change) are either at or under the 50% threshold. Finally, as can be readily seen, this category is reserved for studies that compare treatment and control conditions with a wide range of ways of gathering and examining outcome findings, and this wide range of outcome findings yields a simple majority of significant differences between treatment and control conditions. Further, this category is appropriate for studies that employ limited forms of outcome examination (e.g., limited employment of informants, measures, and/or methods of analysis), and do not yield findings that warrant classification in a category that denotes informant-, measure-, and/or method-specificity in significant differences between treatment and control conditions (please see “Evidence for Contextual- or Informant-Specific Change” and “Evidence for Measure- or Method-Specific Change” category criteria).

Specific criteria for System 1, No Evidence for Change. This criteria is reserved for studies that find absolutely no significant differences between the treatment and control conditions being compared.

Specific criteria for System 1, Evidence for Contextual- or Informant-Specific Change. The “Evidence for Contextual- or Informant-Specific Change” classification coding is reserved for those study comparisons between treatment and control that find significant differences on a grand majority (80%) of findings gathered from specific informant’s ratings. A study is classified in this category when there is a predominantly significant informant-specific pattern of significant findings. A study can be classified in this category if either: **(a)** limited information sources were employed in a study (i.e., less range of outcome informants), and this limited range of outcome examinations yields a pattern of findings that suggests informant-specificity in significant findings; or **(b)** a wide range of information sources were employed in a study, and findings employing a specific informant or informants reveal predominantly significant results.

Again, as in other System 1 categories, a range of findings need to be employed in a study classified in this category. However, this time the range of findings is informant-specific. For example, for a study to be classified in this category based on evidence of parent-specific change, at least three parent-rated findings need to be employed in the study, evidencing significant differences on at least 80% of the findings. Further, limited evidence needs to be found on other findings based on the ratings of other informants. Alternatively, parent-rated findings might have been the only pieces of evidence used in the study; such a study would also be appropriate for this classification.

Note that this category, like all categories, focuses on *differences* between conditions, and not necessarily findings that suggest the treatment condition “outperformed” the control condition. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X		X
Child	Questionnaire	X		
Child	Structured Interview	X	X	
Parent	Questionnaire	–		–
Parent	Questionnaire	X		–
Parent	Questionnaire	–		–
Parent	Structured Interview	–	–	
Parent	Structured Interview	X	X	
Lab Observer	Independent Observation			–
Lab Observer	Independent Observation	–		–
Lab Observer	Independent Observation	X		

Where “X” denotes a significant difference between treatment and control conditions, “–“ denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. In this example, the only predominantly significant results are gleaned from child-rated outcomes, and there are non-significant differences on all parent-rated and lab observer-rated outcomes that are enough to pull every other informant-rated group of outcome findings **UNDER** the simple majority (above 50%) threshold of significant differences. Indeed, each of the other informant-rated outcomes is less than or equal to 50% on significant outcomes. For instance, parent-rated findings are 30% significant, and lab observer-rated findings are 25% significant. Keep in mind that studies can be classified as informant-specific in their patterns of results, and may **ALSO** qualify for inclusion in categories denoting other kinds of specificities of results (e.g., measure and/or method specificity). An example of this kind of study is described below (see “Multiple classifications under System 1” section). Finally, this category is reserved for studies that compare treatment and control conditions either with a limited number or multiple informant-rated outcomes, and yield findings that warrant classification in a category that denotes informant-specificity of significant differences between treatment and control conditions.

Specific criteria for System 1, Evidence for Measure- or Method-Specific Change. The “Evidence for Measure- or Method-Specific Change” classification coding is reserved for those study comparisons between treatment and control that find significant differences on a grand majority (80%) of findings gathered from specific measures of outcome, and/or specific methods of analysis. A study is classified in this category when there is a predominantly significant measure and/or method-specific pattern of significant findings. A study can be classified in this category if either: **(a)** limited measures or methods were employed in a study (i.e., less range of measure or method types), and this limited range of outcome examinations yields a pattern of findings that suggests measure or method-specificity in significant findings; or **(b)** a wide range of measures and/or methods were employed in a study, and findings employing a specific outcome measure type or type of method of analysis reveal predominantly significant results.

Again, as in other System 1 categories, a range of findings need to be employed in a study classified in this category. However, this time the range of findings is measure- and/or method-specific. For example, for a study to be classified in this category based on evidence of change specific to tests of diagnostic status, at least three diagnostic status-based findings need to be employed in the study, evidencing significant differences on at least 80% of the findings. Further, limited evidence needs to be found on other findings based on other methods of examining change. Note that this category, like all categories, focuses on *differences* between conditions, and not necessarily findings that suggest the treatment condition "outperformed" the control condition. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X		–
Child	Questionnaire	X		
Child	Structured Interview	–	–	
Parent	Questionnaire	X		–
Parent	Questionnaire	–		–
Parent	Questionnaire	X		–
Parent	Structured Interview	X	–	
Parent	Structured Interview	X	–	
Lab Observer	Independent Observation	X		–
Lab Observer	Independent Observation	X		–
Lab Observer	Independent Observation	X		–

Where "X" denotes a significant difference between treatment and control conditions, "–" denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. In this example, the only predominantly significant results are gleaned from findings made from comparisons of mean differences between treatment and control conditions, and there are non-significant differences on all findings gleaned from diagnostic status and clinically significant change methods of analysis. Also, findings are sporadically significant on questionnaire, structured interview, and independent observation methods (i.e., at or under 50% significant). Thus, the pattern suggests significant differences between conditions on the mean differences method of analyzing outcomes specifically. Keep in mind that studies can be classified as measure and/or method-specific in their patterns of results, and may **ALSO** qualify for inclusion in the category denoting informant-specificity in significant results. An example of this kind of study is described below (see "Multiple classification codings under System 1" section). Finally, this category is reserved for studies that compare treatment and control conditions either with a limited number or multiple measurement outcome types and/or methods of examining outcomes, and yield findings that warrant classification in a category that denotes measure and/or method-specificity of significant differences between treatment and control conditions.

Multiple classifications under System 1. As mentioned previously, System 1 allows for classification of a single comparison between a treatment condition and a control condition into more than one category. Obviously, a study that compares more than one treatment to one or more control conditions **HAS** to involve more than one classification, because the findings gleaned from each treatment-control comparison that meets criteria for inclusion in the meta-analysis must receive at least one classification under System 1. What this section deals with specifically is when the findings yielded from a single treatment condition compared to a single control condition can be coded into more than one System 1 classification category. A study comparison can be classified into more than one System 1 category only when the findings from this comparison suggest multiple specificities of significant effects, or more than one pattern of informant-, measure-, and/or method-specific significant results. An example here may be helpful. Consider the following table of outcomes:

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	**X**		–
Child	Questionnaire	**X**		–
Child	Questionnaire	**X**		–
Parent	Questionnaire	**X**		–
Parent	Questionnaire	**X**		–
Parent	Questionnaire	**X**		–
Parent	Structured Interview	–	–	
Parent	Structured Interview	–	X	
Lab Observer	Independent Observation	–		X
Lab Observer	Independent Observation	–		X
Lab Observer	Independent Observation	–		

Where “X” denotes a significant difference between treatment and control conditions, “–“ denotes a non-significant difference between treatment and control conditions, and a blank space denotes that an outcome measure was not employed for that particular way of gathering outcome evidence. For ease of presentation, the significant pattern of effects that form the basis of the multiple category classifications are set in bold-type font in the table above. In this example, there are multiple specificities of significant effects. Specifically, predominantly significant effects were found on child- and parent-rated outcomes, assessed using questionnaire ratings, and examined via mean differences examinations between conditions. In this instance, the pattern of informant-, measure-, and method-specific effects suggests a pattern of findings that could be classified in both of the specific effects categories noted previously (see “Evidence for Contextual- or Informant-Specific Change” and “Evidence for Measure- or Method-Specific Change” category criteria). Further, as can be seen in the table, the other ways in which outcomes can be assessed and/or examined are sporadically significant, and either at or under the 50% threshold of significant effects. Please note that in this example, 4 specificities of significant effects were found, but were classified in two System 2 categories. This example is meant to illustrate that many specific effects may be classified either within a single category or between the two System 2 specific-effect categories.

Coding ranges of effect sizes under System 1. After coding the classification category, each classification category code is to be accompanied by a code of the ranges of effect sizes of the findings employed to reach the classification. Ranges should be coded from the smallest effect size gleaned from findings used to reach the classification, to the largest effect size. Effect sizes should be included in the range, regardless of whether the finding from which the effect size was gleaned revealed a significant difference between treatment and control conditions. Effect sizes should be included in the range, regardless of whether they are positive (e.g., treatment “outperformed” control), or negative (control “outperformed” treatment). Further, each effect size range is to be coded for significant effects. That is, for each effect size range, if the lower and/or upper end of the effect size range revealed a significant difference between conditions, this should be marked by an “*” to note the significant difference. Below are examples of coding effect size ranges for different study classifications. Specifically, we present examples of a non-specific study code (i.e., what you would expect from “Best Evidence for Change,” “Evidence for Probable Change,” “Limited Evidence for Change,” and “No Evidence for Change” category code), a single specific study code (i.e., what you would expect from either of the specific-effect categories), and a multiple specific study code (i.e., what you would expect from a classification of more than one informant-, measure-, and/or method-specific pattern of significant effects).

Range of effect sizes for a study classified under “Best Evidence for Change” category.
The table below describes coding ranges of effect sizes for a study comparison classified under the first System 1 category. What will be evident here is that in this category, and like all the other non-specific classification categories (see “Probable Evidence for Change,” “Limited Evidence for Change,” and “No Evidence for Change” category criteria), **ALL** of the findings in the study comparison are used to reach the range of effect sizes. Because in such a study, no specificity of effects is observed, no findings are left out of the effect size range. In this example, .07 is the smallest effect size, and .88 the largest. Only .88 is a significant effect, so it is labeled with the asterisk.

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X (.45)		X (.84)
Child	Questionnaire	X (.62)		
Child	Structured Interview	X (.55)	X (.76)	
Parent	Questionnaire	X (.73)		X (.36)
Parent	Questionnaire	X (.52)		X (.61)
Parent	Questionnaire	– (.12)		X (.43)
Parent	Structured Interview	– (.07)	X (.41)	
Parent	Structured Interview	X (.31)	X (.44)	
Lab Observer	Independent Observation			X (.65)
Lab Observer	Independent Observation	X (.58)		X (.47)
Lab Observer	Independent Observation	X (.88)		
Effect size range: (.07, 88*)				

Range of effect sizes for a study classified under "Evidence for Contextual- or Informant-Specific Change" category. The table below describes coding ranges of effect sizes for a study comparison classified under the first specific-effect System 1 category. What will be evident here is that in this category, and like the other specific-effect classification categories (see "Evidence for Measure- or Method-Specific Change," category criteria), only the findings in the study comparison that are used to reach the category classification are used to reach the range of effect sizes. Thus, only the child-rated findings are included in the effect size range. Because in this study, specificity of effects are observed, findings that are not included in the classification are left out of the effect size range. In this example, .36 is the smallest effect size, and .90 the largest. Both ends of the range revealed a significant effect, so both ends are labeled with the asterisk.

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	X (.45)		X (.64)
Child	Questionnaire	X (.36)		
Child	Structured Interview	X (.75)	X (.90)	
Parent	Questionnaire	– (.08)		– (.14)
Parent	Questionnaire	X (.55)		– (.04)
Parent	Questionnaire	– (.09)		– (.00)
Parent	Structured Interview	– (.11)	– (.06)	
Parent	Structured Interview	X (.98)	X (.66)	
Lab Observer	Independent Observation			– (.10)
Lab Observer	Independent Observation	– (.15)		– (.16)
Lab Observer	Independent Observation	X (.75)		
Effect size range: (.36*, 90*)				

Range of effect sizes for a study classified under both "Evidence for Contextual- or Informant-Specific Change" and "Evidence for Measure- or Method-Specific Change" categories. The table below describes coding ranges of effect sizes for a study comparison classified under multiple specific-effect System 1 categories. Again, only the findings in the study comparison that are used to reach the category classifications are used to reach the range of effect sizes. Thus, only the child- and parent-rated findings, assessed via questionnaire, and examined using mean differences comparisons between treatment and control conditions are included in the effect size range. Because in this study, specificity of effects is observed, findings that are not included in the classification are left out of the effect size range. In this example, .34 is the smallest effect size, and .80 the largest. Both ends of the range revealed a significant effect, so both ends are labeled with the asterisk.

Information Source	Measure Method	Mean Differences	Diagnostic Status	Clinically Significant Change
Child	Questionnaire	**X (.48)**		– (.02)
Child	Questionnaire	**X (.64)**		– (.06)
Child	Questionnaire	**X (.80)**		– (.11)
Parent	Questionnaire	**X (.76)**		– (.14)
Parent	Questionnaire	**X (.34)**		– (.08)
Parent	Questionnaire	**X (.55)**		– (.10)
Parent	Structured Interview	– (.12)	– (.09)	
Parent	Structured Interview	– (.00)	X (.48)	
Lab Observer	Independent Observation	– (.04)		X (.52)
Lab Observer	Independent Observation	– (.03)		X (.66)
Lab Observer	Independent Observation	– (.07)		
Effect size range: (.34*, 80*)				

Classification of studies under System 2. Codes for classifications of studies under System 2 will be based on the system noted in the tables below. As mentioned previously, information used to classify studies based on this system has already been collected in prior sections. This code compiles this information to arrive at a single conclusion as to what the evidence for a study says about the treatment that was examined in it. Unlike the previous system, however, only a single conclusion, and thus, a single code ought to be made for each eligible treatment condition studied in the article, no matter how many control conditions it was compared against, or how many outcome measures of the primary problem targeted in treatment were examined. This does not mean that a single study can only be classified once under System 2; if two different treatments in a study that each meets criteria for inclusion in the meta-analysis are compared to one or more control conditions, each treatment comparison receives a single code. However, if a treatment in a study is compared to more than one control condition, that treatment may only be classified once (e.g., under the System 2, Table 2 criteria, a treatment has to outperform **A** no-treatment control condition in a study, not **ALL** no-treatment control conditions in a study). **What should be clear in classifying studies under System 2 is that prior classifications for a study made under System 1 should not be taken into account when classifying the same study under System 2.** Please use System 2 criteria and only System 2 criteria to classify a given study under System 2. Two tables outlining the different classification codes that may be made under System 2 are provided below, as well as a more detailed description of each System 2 classification code.

What will appear obvious from the tables below is that they are not established to classify individual studies per se, but treatments themselves. Thus, with System 2, please code individual studies in terms of whether they should be considered or counted as a study that meets criteria for the categories in either Table 1 or Table 2. Perhaps most importantly, please only use evidence or data from treatment comparisons coded previously. That is, only comparisons between treatment and control conditions that meet criteria for inclusion in the meta-analysis should be used to code classifications in System 2 categories. **Further, please only use consensus codes on significant differences between conditions to make System 2 classifications (i.e., information coded for Parts 1-3, and not information from the article itself).** Criteria for each category in Table 1 and Table 2 are further described below after each System 2 category table.

NOTE ON THE LAST TWO CRITERIA FOR TABLES 1 AND 2 (MANUAL CRITERIA AND SAMPLE CHARACTERISTICS). EACH OF THE TABLE CRITERIA IN SYSTEM 2 REQUIRES THAT STUDIES USE TREATMENT MANUALS FOR THE TREATMENT EXAMINED, AND PROVIDE SAMPLE CHARACTERISTICS. TYPICALLY, YOU CAN TELL THAT A TREATMENT MANUAL WAS USED IN A STUDY IN A NUMBER OF WAYS: (A) IF THE AUTHORS CITE A PUBLICATION THAT DESCRIBES THE INTERVENTION; (B) A SESSION-BY-SESSION DESCRIPTION OF THE INTERVENTION IS PROVIDED IN THE ARTICLE; OR (C) SOME DESCRIPTION OF THE TREATMENT IS PROVIDED, AND THE DESCRIPTION IS LABLED AS A “PROTOCOL.” ESSENTIALLY, THE MANUAL CRITERIA IS SATISFIED IF THE AUTHORS PROVIDE ANY DESCRIPTION OF THE INTERVENTION THAT GIVES YOU AN IDEA THAT THE AUTHORS ADMINISTERED AN INTERVENTION WITH SOME SORT OF STRUCUTURE, AND THAT STRUCTURE WAS CODIFIED IN SOME WAY, SO THAT THERAPISTS KNEW WHAT THE INTERVENTION WAS SUPPOSED TO BE, AND THERE WERE AT LEAST SOME GENERAL GUIDELINES OF HOW TO ADMINISTER THE INTERVENTION. KEEP IN MIND THAT THE INTERVENTION’S DESCRIPTION DOES NOT HAVE TO BE PUBLISHED FOR THE INTERVENTION TO BE CONSIDERED MANUALIZED.

WITH REGARD TO SAMPLE CHARACTERISTICS, YOU PROBABLY HAVE A GOOD IDEA OF WHETHER THE AUTHORS CLEARLY SPECIFIED THE SAMPLE CHARACTERISTICS IF THEY DESCRIBE WHERE AND HOW THE SAMPLE WAS RECRUITED, THE SIZE OF THE SAMPLE, AND WHAT THE SEVERITY OF THE SAMPLE’S PROBLEMS WERE PRIOR TO TREATMENT (I.E., PRE-TREATMENT RATINGS OF THE SAMPLE ON MEASURES THAT ARE TO BE EMPLOYED TO EXAMINE OUTCOMES).

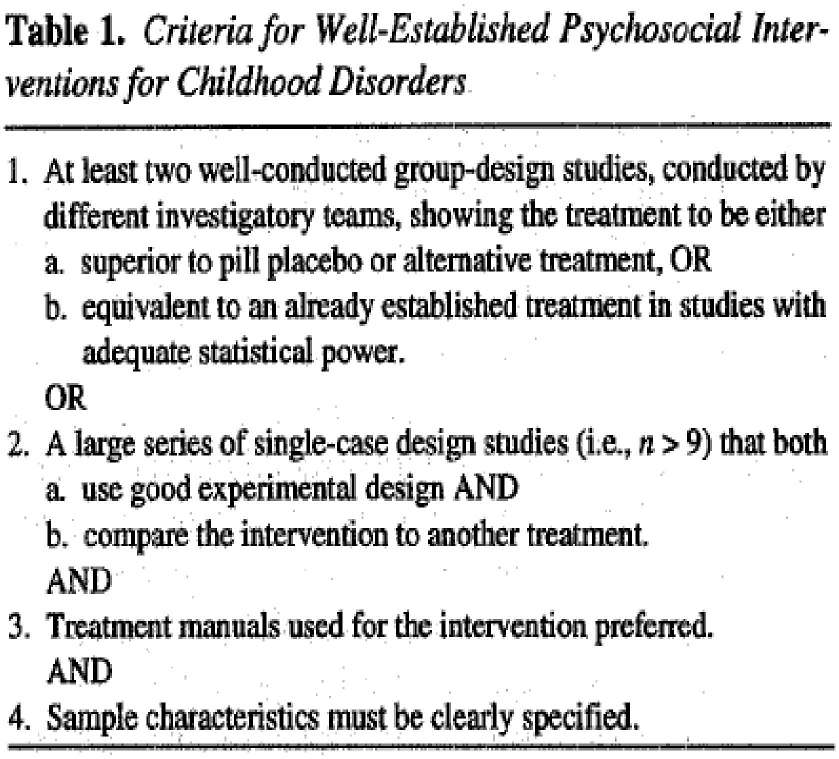


Table 1. *Criteria for Well-Established Psychosocial Interventions for Childhood Disorders*

1. At least two well-conducted group-design studies, conducted by different investigatory teams, showing the treatment to be either
   a. superior to pill placebo or alternative treatment, OR
   b. equivalent to an already established treatment in studies with adequate statistical power.

   OR
2. A large series of single-case design studies (i.e., $n > 9$) that both
   a. use good experimental design AND
   b. compare the intervention to another treatment.

   AND
3. Treatment manuals used for the intervention preferred.

   AND
4. Sample characteristics must be clearly specified.

Specific criteria for System 2, Table 1. For Table 1, it is important to reiterate that only treatment-control comparisons that were coded in the meta-analysis are to be used to decide whether a study meets criteria for consideration in the System 2, Table 1 category. As with System 1 category criteria, data coded in Parts 1-3 should be utilized to reach conclusions on codes for System 2 classification categories.

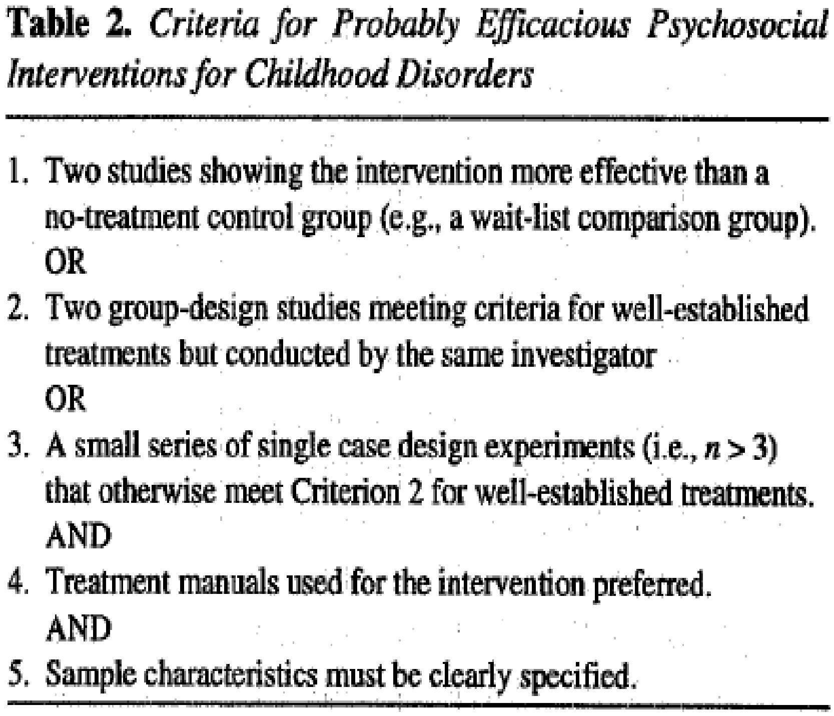


Table 2. *Criteria for Probably Efficacious Psychosocial Interventions for Childhood Disorders*

1. Two studies showing the intervention more effective than a no-treatment control group (e.g., a wait-list comparison group).

   OR
2. Two group-design studies meeting criteria for well-established treatments but conducted by the same investigator

   OR
3. A small series of single case design experiments (i.e., $n > 3$) that otherwise meet Criterion 2 for well-established treatments.

   AND
4. Treatment manuals used for the intervention preferred.

   AND
5. Sample characteristics must be clearly specified.

Specific criteria for System 2, Table 2. For Table 2, it is again important to reiterate that only treatment-control comparisons that were coded in the meta-analysis are to be used to decide whether a study meets criteria for consideration in the System 2, Table 2 category. As with System 1 category criteria, data coded in Parts 1-3 should be utilized to reach conclusions on codes for System 2 classification categories.

Data Collection Part 4 (cont'd): Coding Classifications of Studies

Now that System 1 and System 2 category coding criteria have been described and illustrated, descriptions of specific study codes will now be provided. The coding sheet is structured to allow for coding multiple study comparisons (i.e., studies that compare multiple treatments to one or more control conditions). Items 1a and 1b have already been completed for you, as they were transferred from Part 1 of consensus codes.

(1) *Classification of Studies of CBT for Child Anxiety Problems.* The first set of codes in Part 4 regard the classification of each treatment-control comparison included in the meta-analysis for CBT for child anxiety problems (Items 1a-1g). That is, each CBT treatment-control comparison is to be classified under System 1, and each treatment-control comparison that is included in the meta-analysis is to be classified once under System 2, even if that treatment is compared to more than one control condition within a study that meets criteria for inclusion in the meta-analysis. Descriptions of each code in Item 1 are provided below.

(1a) *Author name and year.* Simply transfer from Items 1 and 2 from Part 1 of the coding sheet.

(1b) *Groups compared?* Simply transfer from Items 4 and 5 from Part 1 of the coding sheet.

(1c) *Manual (Y/N)?* Please code "Yes" or "No" on whether there was a manual in the treatment-control comparison being coded. As mentioned in the previous note on manual criteria, there are a number of ways in which you can tell whether a manual was utilized for the treatment (please consult note above for specific details). Essentially, if the study provides information that leads you to believe that the treatment administered was administered with some codified structure in mind (e.g., published manual or description of the intervention, session-by-session protocol or general outline of intervention, standard videotape administered to subjects forms the basis of intervention), a manual was probably used to administer the treatment. You can usually find this information in the Method section, under a subsection describing the treatment.

(1d) *Manual Citation.* Please code the citation of the manual provided by the authors. If no citation is provided, please code "unnamed manual" for Item 1d.

(1e) *System 1 Classification Category.* Please follow the criteria described above to code a System 1 classification category for the treatment-control study comparison. Multiple classification codes can be employed if the study comparison meets criteria for both categories, or if more than one pattern of specific significant effects meet criteria for a single code (e.g., Code 5 can be employed if informant-specific significant effects are found for both parent- and child-rated outcomes). Codes should be made on the first line of Item 1e using the following:

(1) Best Evidence for Change
(2) Evidence for Probable Change
(3) Limited Evidence for Change
(4) No Evidence for Change
(5) Evidence for Contextual- or Informant-Specific Change
(6) Evidence for Measure- or Method-Specific Change

For Codes 5 and 6, please specify the pattern of significant effects on the second line of Item 1e on the coding sheet.

(1f) *System 1 Effect Size Range with Category.* Please follow the directions described above to code a System 1 effect size range for the System 1 classification category coded in Item 1e. Again, effect size ranges should only include findings that are employed to make the System 1 classification category code(s) made in Item 1e.

(1g) *System 2 Classification Category.* Please follow the criteria described above to code a System 2 classification category for the treatment-control study comparison. Again, only a single conclusion, and thus, a single code ought to be made for each eligible treatment studied in the article, no matter how many control conditions it was compared against, or how many outcome measures of the primary problem targeted in treatment were examined. This does not mean that a single study can only be classified once under System 2; if two treatments in a study that each meet criteria for inclusion in the meta-analysis are compared to a control condition, each treatment comparison may receive a single code. However, if a treatment in a study is compared to more than one control condition, that treatment may only be classified once. Codes should be made on the first line of Item 1g using the following:

(1) System 2, Table 1
(2) System 2, Table 2
(3) No code (study did not meet criteria for either System 2 category)

(2) *Summary of Classifications of Studies of CBT for Child Anxiety Problems Using System 1 (Item 2a through Item 2e).* In Item 1 in this section, you classified findings gleaned from every treatment-control comparison that met criteria for inclusion in this meta-analysis. In Item 2 of this section, you will summarize the codes made in Item 1. The first summary (Item 2a through Item 2e) will consist of compiling the number of codes you made for each classification category in System 1, and tallying them in a "box score" across all of the System 1 categories. The tallies you will make across the System 1 classification categories will consist of two types: **(a)** a tally of the total amount of treatment-control study comparisons coded in Item 1; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol.

Note about Manuals. Again, keep in mind that two studies could use the same manual or protocol, even though that manual or protocol was unpublished. Also, deciding on whether the same manual was used for the same treatment across studies could be tricky, because often times an author may adapt their manual or protocol from that of another author's manual or protocol. In those cases, you can be reasonably certain that the same manual or protocol was used if one of the authors noted that their manual or protocol was an adaptation of the other author's manual or protocol. Another case where you can be reasonably certain that the same manual was used is if an author notes that their study is a replication and extension of a previous clinical trial that they or another author conducted with the same treatment. In all cases, please make sure that manuals are administered in the same format for the two or more study comparisons that you code as having the same manual. That is, for instance if a manual is administered in an individual format in one study, the other study that used the same or adapted manual must also administer their treatment in an individual format for those two studies to be counted as administering the same treatment with the same manual.

(2a) *Summarizing System 1 Classification Categories: Total Studies.*

(2b) *Summarizing System 1 Classification Categories: Manual 1.*

(2c) *Summarizing System 1 Classification Categories: Manual 2.*

(2d) *Summarizing System 1 Classification Categories: Manual 3.*

(2e) *Summarizing System 1 Classification Categories: Manual 4.*

(2) *Summary of Classifications of Studies of CBT for Child Anxiety Problems Using System 1 (Item 2f through Item 2j).* The second summary (Item 2f through Item 2j) will consist of three steps: **(1)** compiling the codes you made for the ranges of effect sizes associated with each classification category code made under System 1; **(2)** classifying these effect size ranges using the effect size conventions of "small," "medium," and "large" effects advanced by Cohen (1988); and **(3)** tallying these classifications of the effect size ranges in another box score across all of the possible permutations of effect size ranges under Cohen (1988). As you may recall from your research methods class, Cohen advanced effect size conventions for three kinds of effect sizes: **(a)** small effects (.20); **(b)** medium effects (.50); and **(c)** large effects (.80). This coding system will add another convention: below small (anything under .20, including negative effect sizes, where the control condition "outperforms" the treatment condition). The following categories are employed in this effect size range summary system:

(1) "Below Small" to "Below Small": (Lower end includes any effect size below .20;
Upper end includes any effect size below .20)
(2) "Below Small" to "Small": (Lower end includes any effect size below .20;
Upper end includes any effect size greater than or equal to .20, but less than .50)
(3) "Below Small" to "Medium": (Lower end includes any effect size below .20;
Upper end includes any effect size greater than or equal to .50, but less than .80)
(4) "Below Small" to "Large": (Lower end includes any effect size below .20;
.Upper end includes any effect size greater than or equal to .80)
(5) "Small" to "Small": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .20, but less than .50)
(6) "Small" to "Medium": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .50, but less than .80)
(7) "Small" to "Large": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .80)
(8) "Medium" to "Medium": (Lower end includes any effect size greater than or equal to .50,
but less than .80; Upper end includes any effect size greater than or equal to .50, but less than .80)
(9) "Medium" to "Large": (Lower end includes any effect size greater than or equal to .50,
but less than .80; Upper end includes any effect size greater than or equal to .80)
(10) "Large" to "Large": (Lower end includes any effect size greater than or equal to .80;
Upper end includes any effect size greater than or equal to .80)

Again, the tallies you will make across the effect size range classifications consists of two types: **(a)** a tally of the total amount of treatment-control study comparisons; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 2a through Item 2e). For each tally, please code the number of treatment-control study comparisons that fit into each of the effect size range conventions.

(2f) *Summarizing System 1 Ranges of Effect Sizes: Total Studies.*

(2g) *Summarizing System 1 Ranges of Effect Sizes: Manual 1.*

(2h) *Summarizing System 1 Ranges of Effect Sizes: Manual 2.*

(2i) *Summarizing System 1 Ranges of Effect Sizes: Manual 3.*

(2j) *Summarizing System 1 Ranges of Effect Sizes: Manual 4.*

(2) *Summary of Classifications of Studies of CBT for Child Anxiety Problems Using System 1 (Item 2k through Item 2p).* In the third summary (Item 2k through Item 2p), we will consolidate all of the information from the two prior summaries, and look for studies that were classified in the same System 1 category (or in the case of multiple classifications, multiple categories). First, you will look for studies that were classified in the same System 1 category or categories. By "same category," we mean at least two studies that share the exact same classification of study findings. In the case of the specific-effects categories (i.e., "Evidence for Contextual- or Informant-Specific Change," "Evidence for Measure- or Method-Specific Change"), the classification must be for the exact same pattern of specificity of effects. For instance, if two studies are classified in the informant-specific category, and one study found informant-specific effects for parent-rated findings, and the other study found informant-specific effects for child-rated findings, these two studies **DID NOT** make findings that could be classified in the same categories. The same goes for studies that were classified in multiple specific-effects categories, the **EXACT** classification of specific effects must be made across the studies for them to be coded as being classified in the same System 1 category or categories. However, please keep in mind that studies that were classified in non-specific System 1 categories ("Best Evidence for Change," "Probable Evidence for Change," "Limited Evidence for Change," "No Evidence for Change") **DO NOT** have to employ the exact same informants, measures, and methods in their studies to be classified in the same category. Stated another way, two studies classified in the "Probable Evidence for Change" category can be coded as being classified in the same System 1 category, regardless of the fact that Study 1 employed different informants, measures, and methods, relative to Study 2.

Second, once studies that can be classified in the same System 1 classification category or categories are identified, it is important to decipher whether those studies **ALSO** exhibited effect size ranges that could be classified in the same effect size convention ranges. This second comparison takes into account the second summary performed in this section, based on Cohen (1988) effect size conventions. Studies must be classified in the same System 1 category or categories, and exhibit the same effect size convention ranges for these studies to reach the final step of the section.

Finally, for studies that pass through both of these stages (i.e., more than one study that could be classified in the same System 1 category, and exhibit the same effect size ranges, as classified using Cohen [1988]), please write down 4 pieces of information: **(a)** the citations of these studies; **(b)** the treatment and control groups being compared within the similar classifications and effect size ranges being coded; **(c)** the System 1 classifications of these studies; and **(d)** their effect size ranges and effect size convention classifications. If more than one similarity is coded, please note on the coding sheet which studies were similar to each other.

Again, the coding of studies that fall within the same System 1 classification categories and effect size ranges consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 2a through Item 2e).

(2k) *Classifications of Studies in the Same System 1 Category: Total Studies.*

(2l) *Classifications of Studies in the Same System 1 Category: Manual 1.*

(2m) *Classifications of Studies in the Same System 1 Category: Manual 2.*

(2n) *Classifications of Studies in the Same System 1 Category: Manual 3.*

(2o) *Classifications of Studies in the Same System 1 Category: Manual 4.*

(2p) *Classifications of Studies in the Same System 1 Category: Manual 5.*

(3) *Summary of Classifications of Studies of CBT for Child Anxiety Problems Using System 2 (Item 3a through Item 3i).* In Item 1 in this section, you classified findings gleaned from every treatment-control comparison that met criteria for inclusion in this meta-analysis. In Item 2, you summarized codes made of treatment-control comparisons under System 1. In Item 3 of this section, you will summarize the codes made of treatment-control comparisons under System 2. The first summary made in Item 3 (Item 3a through Item 3i) will consist of compiling the number of codes you made for each classification category in System 2, and tallying them in a "box score" across both of the System 2 categories. The tallies you will make across the System 2 categories will consist of two types: **(a)** a tally of the total amount of treatment-control study comparisons coded in Item 1; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol.

Note about Manuals. Again, keep in mind that two studies could use the same manual or protocol, even though that manual or protocol was unpublished. Also, deciding on whether the same manual was used for the same treatment across studies could be tricky, because often times an author may adapt their manual or protocol from that of another author's manual or protocol. In those cases, you can be reasonably certain that the same manual or protocol was used if one of the authors noted that their manual or protocol was an adaptation of the other author's manual or protocol. Another case where you can be reasonably certain that the same manual was used is if an author notes that their study is a replication and extension of a previous clinical trial that they or another author conducted with the same treatment. In all cases, please make sure that manuals are administered in the same format for the two or more study comparisons that you code as having the same manual. That is, for instance if a manual is administered in an individual format in one study, the other study that used the same or adapted manual must also administer their treatment in an individual format for those two studies to be counted as administering the same treatment with the same manual.

(3a) *Summarizing System 2 Classification Categories: Total Studies.*

(3b) *Summarizing System 2 Classification Categories: Manual 1.*

(3c) *Summarizing System 2 Classification Categories: Manual 2.*

(3d) *Summarizing System 2 Classification Categories: Manual 3.*

(3e) *Summarizing System 2 Classification Categories: Manual 4.*

(3f) *Summarizing System 2 Classification Categories: Manual 5.*

(3g) *Summarizing System 2 Classification Categories: Manual 6.*

(3h) *Summarizing System 2 Classification Categories: Manual 7.*

(3i) *Summarizing System 2 Classification Categories: Manual 8.*

(3) *Summary of Classifications of Studies of CBT for Child Anxiety Problems Using System 2 (Item 3j through Item 3o).* In the second summary of Item 3 summary (Item 3j through Item 3o), we will consolidate all of the information from the prior summary, and look for studies that were classified in the same System 2 category. First, you will look for studies that were classified in the same System 2 category. By "same category," we mean at least two studies that share the exact same System 2 classification. For instance, if two studies are classified in the System 2, Table 2 category, these two studies **DID** make findings that could be classified in the same category.

Second, for studies that pass through the first stage, please write down 3 pieces of information: **(a)** the citations of these studies; **(b)** the treatment and control groups being compared within the similar classifications being coded; and **(c)** the System 2 classifications of these studies. If more than one similarity is coded, please note on the coding sheet which studies were similar to each other.

Again, the coding of studies that fall within the same System 2 classification categories consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 3a through Item 3i).

(3j) *Classifications of Studies in the Same System 2 Categories: Total Studies.*

(3k) *Classifications of Studies in the Same System 2 Categories: Manual 1.*

(3l) *Classifications of Studies in the Same System 2 Categories: Manual 2.*

(3m) *Classifications of Studies in the Same System 2 Categories: Manual 3.*

(3n) *Classifications of Studies in the Same System 2 Categories: Manual 4.*

(3o) *Classifications of Studies in the Same System 2 Categories: Manual 5.*

(4) *Classification of Studies of BPT for Child Conduct Problems.* The first set of codes in Part 4 regard the classification of each treatment-control comparison included in the meta-analysis for BPT for child conduct problems (Items 4a-4g). That is, each BPT treatment-control comparison is to be classified under System 1, and each treatment-control comparison that is included in the meta-analysis is to be classified once under System 2, even if that treatment is compared to more than one control condition within a study that meets criteria for inclusion in the meta-analysis. Descriptions of each code in Item 4 are provided below. Items 4a and 4b have already been completed for you, as they were transferred from Part 1 of consensus codes.

(4a) *Author name and year*. Simply transfer from Items 1 and 2 from Part 1 of the coding sheet.

(4b) *Groups compared?* Simply transfer from Items 4 and 5 from Part 1 of the coding sheet.

(4c) *Manual (Y/N)?* Please code "Yes" or "No" on whether there was a manual in the treatment-control comparison being coded. As mentioned in the previous note on manual criteria, there are a number of ways in which you can tell whether or not a manual was utilized for the treatment (please consult note above for specific details). Essentially, if the study provides information that leads you to believe that the treatment administered was administered with some codified structure in mind (e.g., published manual or description of the intervention, session-by-session protocol or general outline of intervention, standard videotape administered to subjects forms the basis of intervention), a manual was probably used to administer the treatment. You can usually find this information in the Method section, under a subsection describing the treatment.

(4d) *Manual Citation.* Please code the citation of the manual provided by the authors. If no citation is provided, please code "unnamed manual" for Item 1d.

(4e) *System 1 Classification Category.* Please follow the criteria described above to code a System 1 classification category for the treatment-control study comparison. Multiple classification codes can be employed if the study comparison meets criteria for both categories, or if more than one pattern of specific significant effects meet criteria for a single code (e.g., Code 5 can be employed if informant-specific significant effects are found for both parent- and child-rated outcomes). Codes should be made on the first line of Item 4e using the following:

(1) Best Evidence for Change
(2) Evidence for Probable Change
(3) Limited Evidence for Change
(4) No Evidence for Change
(5) Evidence for Contextual- or Informant-Specific Change
(6) Evidence for Measure- or Method-Specific Change

For Codes 5 and 6, please specify the pattern of significant effects on the second line of Item 4e on the coding sheet.

(4f) *System 1 Effect Size Range with Category.* Please follow the directions described above to code a System 1 effect size range for the System 1 classification category coded in Item 4e. Again, effect size ranges should only include findings that are employed to make the System 1 classification category code(s) made in Item 4e.

(4g) *System 2 Classification Category.* Please follow the criteria described above to code a System 2 classification category for the treatment-control study comparison. Again, only a single conclusion, and thus, a single code ought to be made for each eligible treatment studied in the article, no matter how many control conditions it was compared against, or how many outcome measures of the primary problem targeted in treatment were examined. This does not mean that a single study can only be classified once under System 2; if two treatments in a study that each meet criteria for inclusion in the meta-analysis are compared to a control condition, each treatment comparison may receive a single code. However, if a treatment in a study is compared to more than one control condition, that treatment may only be classified once. Codes should be made on the first line of Item 4g using the following:

(1) System 2, Table 1
(2) System 2, Table 2
(3) No code (study did not meet criteria for either System 2 category)

(5) *Summary of Classifications of Studies of BPT for Child Conduct Problems Using System 1 (Item 5a through Item 5e).* In Item 4 in this section, you classified findings gleaned from every treatment-control comparison that met criteria for inclusion in this meta-analysis. In Item 5 of this section, you will summarize the codes made in Item 4. The first summary (Item 5a through Item 5e) will consist of compiling the number of codes you made for each classification category in System 1, and tallying them in a "box score" across all of the System 1 categories. The tallies you will make across the System 1 classification categories will consist of two types: **(a)** a tally of the total amount of treatment-control study comparisons coded in Item 4; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol.

Note about Manuals. Again, keep in mind that two studies could use the same manual or protocol, even though that manual or protocol was unpublished. Also, deciding on whether the same manual was used for the same treatment across studies could be tricky, because often times an author may adapt their manual or protocol from that of another author's manual or protocol. In those cases, you can be reasonably certain that the same manual or protocol was used if one of the authors noted that their manual or protocol was an adaptation of the other author's manual or protocol. Another case where you can be reasonably certain that the same manual was used is if an author notes that their study is a replication and extension of a previous clinical trial that they or another author conducted with the same treatment. In all cases, please make sure that manuals are administered in the same format for the two or more study comparisons that you code as having the same manual. That is, for instance if a manual is administered in an individual format in one study, the other study that used the same or adapted manual must also administer their treatment in an individual format for those two studies to be counted as administering the same treatment with the same manual.

(5a) *Summarizing System 1 Classification Categories: Total Studies.*

(5b) *Summarizing System 1 Classification Categories: Manual 1.*

(5c) *Summarizing System 1 Classification Categories: Manual 2.*

(5d) *Summarizing System 1 Classification Categories: Manual 3.*

(5e) *Summarizing System 1 Classification Categories: Manual 4.*

(5) *Summary of Classifications of Studies of BPT for Child Conduct Problems Using System 1 (Item 5f through Item 5j).* The second summary (Item 5f through Item 5j) will consist of three steps: **(1)** compiling the codes you made for the ranges of effect sizes associated with each classification category code made under System 1; **(2)** classifying these effect size ranges using the effect size conventions of "small," "medium," and "large" effects advanced by Cohen (1988); and **(3)** tallying these classifications of the effect size ranges in another box score across all of the possible permutations of effect size ranges under Cohen (1988). As you may recall from your research methods class, Cohen advanced effect size conventions for three kinds of effect sizes: **(a)** small effects (.20); **(b)** medium effects (.50); and **(c)** large effects (.80). This coding system will add another convention: below small (anything under .20, including negative effect sizes, where the control condition "outperforms" the treatment condition). The following categories are employed in this effect size range summary system:

(1) "Below Small" to "Below Small": (Lower end includes any effect size below .20;
Upper end includes any effect size below .20)
(2) "Below Small" to "Small": (Lower end includes any effect size below .20;
Upper end includes any effect size greater than or equal to .20, but less than .50)
(3) "Below Small" to "Medium": (Lower end includes any effect size below .20;
Upper end includes any effect size greater than or equal to .50, but less than .80)
(4) "Below Small" to "Large": (Lower end includes any effect size below .20;
.Upper end includes any effect size greater than or equal to .80)
(5) "Small" to "Small": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .20, but less than .50)
(6) "Small" to "Medium": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .50, but less than .80)
(7) "Small" to "Large": (Lower end includes any effect size greater than or equal to .20,
but less than .50; Upper end includes any effect size greater than or equal to .80)
(8) "Medium" to "Medium": (Lower end includes any effect size greater than or equal to .50,
but less than .80; Upper end includes any effect size greater than or equal to .50, but less than .80)
(9) "Medium" to "Large": (Lower end includes any effect size greater than or equal to .50,
but less than .80; Upper end includes any effect size greater than or equal to .80)
(10) "Large" to "Large": (Lower end includes any effect size greater than or equal to .80;
Upper end includes any effect size greater than or equal to .80)

Again, the tallies you will make across the effect size range classifications will consist of two types: **(a)** a tally of the total amount of treatment-control study comparisons coded in Item 4; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 5a through Item 5f). For each tally, please code the number of treatment-control study comparisons that fit into each of the effect size range conventions.

(5f) *Summarizing System 1 Ranges of Effect Sizes: Total Studies.*

(5g) *Summarizing System 1 Ranges of Effect Sizes: Manual 1.*

(5h) *Summarizing System 1 Ranges of Effect Sizes: Manual 2.*

(5i) *Summarizing System 1 Ranges of Effect Sizes: Manual 3.*

(5j) *Summarizing System 1 Ranges of Effect Sizes: Manual 4.*

(5) *Summary of Classifications of Studies of BPT for Child Conduct Problems Using System 1 (Item 5k through Item 5p).* In the third summary (Item 5k through Item 5p), we will consolidate all of the information from the two prior summaries, and look for studies that were classified in the same System 1 category (or in the case of multiple classifications, multiple categories). First, you will look for studies that were classified in the same System 1 category or categories. By “same category,” we mean at least two studies that share the exact same classification of study findings. In the case of the specific-effects categories (i.e., “Evidence for Contextual- or Informant-Specific Change,” “Evidence for Measure- or Method-Specific Change”), the classification must be for the exact same pattern of specificity of effects. For instance, if two studies are classified in the informant-specific category, and one study found informant-specific effects for parent-rated findings, and the other study found informant-specific effects for child-rated findings, these two studies **DID NOT** make findings that could be classified in the same categories. The same goes for studies that were classified in multiple specific-effects categories, the **EXACT** classification of specific effects must be made across the studies for them to be coded as being classified in the same System 1 category or categories. However, please keep in mind that studies that were classified in non-specific System 1 categories (“Best Evidence for Change,” “Probable Evidence for Change,” “Limited Evidence for Change,” “No Evidence for Change”) **DO NOT** have to employ the exact same informants, measures, and methods in their studies to be classified in the same category. Stated another way, two studies classified in the “Probable Evidence for Change” category can be coded as being classified in the same System 1 category, regardless of the fact that Study 1 employed different informants, measures, and methods, relative to Study 2.

Second, once studies that can be classified in the same System 1 classification category or categories are identified, it is important to decipher whether those studies **ALSO** exhibited effect size ranges that could be classified in the same effect size convention ranges. This second comparison takes into account the second summary performed in this section, based on Cohen (1988) effect size conventions. Studies must be classified in the same System 1 category or categories, and exhibit the same effect size convention ranges for these studies to reach the final step of the section.

Finally, for studies that pass through both of these stages (i.e., more than one study that could be classified in the same System 1 category, and exhibit the same effect size ranges, as classified using Cohen [1988]), please write down 4 pieces of information: **(a)** the citations of these studies; **(b)** the treatment and control groups being compared within the similar classifications and effect size ranges being coded; **(c)** the System 1 classifications of these studies; and **(d)** their effect size ranges and effect size convention classifications. If more than one similarity is coded, please note on the coding sheet which studies were similar to each other.

Again, the coding of studies that fall within the same System 1 classification categories and effect size ranges consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 5a through Item 5f).

(5k) *Classifications of Studies in the Same System 1 Category: Total Studies.*

(5l) *Classifications of Studies in the Same System 1 Category: Manual 1.*

(5m) *Classifications of Studies in the Same System 1 Category: Manual 2.*

(5n) *Classifications of Studies in the Same System 1 Category: Manual 3.*

(5o) *Classifications of Studies in the Same System 1 Category: Manual 4.*

(5p) *Classifications of Studies in the Same System 1 Category: Manual 5.*

(6) *Summary of Classifications of Studies of BPT for Child Conduct Problems Using System 2 (Item 6a through Item 6i).* In Item 4 in this section, you classified findings gleaned from every treatment-control comparison that met criteria for inclusion in this meta-analysis. In Item 5, you summarized codes made of treatment-control comparisons under System 1. In Item 6 of this section, you will summarize the codes made of treatment-control comparisons under System 2. The first summary made in Item 6 (Item 6a through Item 6i) will consist of compiling the number of codes you made for each classification category in System 2, and tallying them in a "box score" across both of the System 2 categories. The tallies you will make across the System 2 categories will consist of two types: **(a)** a tally of the total amount of treatment-control study comparisons coded in Item 4; **(b)** tallies of studies that examined the outcomes of a treatment administered using the same manual or treatment protocol.

Note about Manuals. Again, keep in mind that two studies could use the same manual or protocol, even though that manual or protocol was unpublished. Also, deciding on whether the same manual was used for the same treatment across studies could be tricky, because often times an author may adapt their manual or protocol from that of another author's manual or protocol. In those cases, you can be reasonably certain that the same manual or protocol was used if one of the authors noted that their manual or protocol was an adaptation of the other author's manual or protocol. Another case where you can be reasonably certain that the same manual was used is if an author notes that their study is a replication and extension of a previous clinical trial that they or another author conducted with the same treatment. In all cases, please make sure that manuals are administered in the same format for the two or more study comparisons that you code as having the same manual. That is, for instance if a manual is administered in an individual format in one study, the other study that used the same or adapted manual must also administer their treatment in an individual format for those two studies to be counted as administering the same treatment with the same manual.

(6a) *Summarizing System 2 Classification Categories: Total Studies.*

(6b) *Summarizing System 2 Classification Categories: Manual 1.*

(6c) *Summarizing System 2 Classification Categories: Manual 2.*

(6d) *Summarizing System 2 Classification Categories: Manual 3.*

(6e) *Summarizing System 2 Classification Categories: Manual 4.*

(6f) *Summarizing System 2 Classification Categories: Manual 5.*

(6g) *Summarizing System 2 Classification Categories: Manual 6.*

(6h) *Summarizing System 2 Classification Categories: Manual 7.*

(6i) *Summarizing System 2 Classification Categories: Manual 8.*

(6) *Summary of Classifications of Studies of BPT for Child Conduct Problems Using System 2 (Item 6j through Item 6o).* In the second summary of Item 6 summary (Item 6j through Item 6o), we will consolidate all of the information from the prior summary, and look for studies that were classified in the same System 2 category. First, you will look for studies that were classified in the same System 2 category. By "same category," we mean at least two studies that share the exact same System 2 classification. For instance, if two studies are classified in the System 2, Table 2 category, these two studies **DID** make findings that could be classified in the same category.

Second, for studies that pass through the first stage, please write down 3 pieces of information: **(a)** the citations of these studies; **(b)** the treatment and control groups being compared within the similar classifications being coded; and **(c)** the System 2 classifications of these studies. If more than one similarity is coded, please note on the coding sheet which studies were similar to each other.

Again, the coding of studies that fall within the same System 2 classification categories consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous note on manuals (Item 6a through Item 6i).

(6j) *Classifications of Studies in the Same System 2 Categories: Total Studies.*

(6k) *Classifications of Studies in the Same System 2 Categories: Manual 1.*

(6l) *Classifications of Studies in the Same System 2 Categories: Manual 2.*

(6m) *Classifications of Studies in the Same System 2 Categories: Manual 3.*

(6n) *Classifications of Studies in the Same System 2 Categories: Manual 4.*

(6o) *Classifications of Studies in the Same System 2 Categories: Manual 5.*

IX. Data Collection Part 5: Coding Classifications of Treatments

Thank you very much for coding the studies thus far! In the last section, we will be taking the information coded in Part 4 to answer some questions pertaining to classifying the evidence for the treatments being studied in this meta-analysis. Thus, in Part 5 the emphasis shifts away from classifying studies of CBT and BPT to classifying the CBT and BPT treatments themselves, based on the evidence gathered in the meta-analysis. First, we will be consolidating information gathered from coding studies of CBT and BPT under System 1 to answer Yes/No questions about whether CBT interventions change or reduce child anxiety problems, and whether BPT interventions change or reduce child conduct problems. Second, we will be consolidating information gathered from coding studies of CBT and BPT under System 2 to answer Yes/No questions about whether CBT and BPT interventions fall into either of the two System 2 categories (System 2, Table 1 criteria; System 2, Table 2 criteria). Similar to the last few items of Part 4, codes made in Part 5 take into account the totality of information coded across all studies included in the meta-analysis. Thus, Parts 4 and 5 should be coded only after all of the information from Parts 1 through 3 is coded for all studies. Similar to Part 4, Yes/No answers to the questions below will be made across all studies of CBT and all studies of BPT, and then also within pools of studies of the same treatment, using the same manual. As always, please contact me with any questions that you may have regarding study codes in Part 5.

(1) *Classifications of Treatments Using System 1, CBT for Child Anxiety Problems (Item 1a through Item 1x).* In this first section of Yes/No questions for CBT for child anxiety problems, we will answer Yes/No questions, based on codes we made in Part 4, under System 1. All of the questions from Item 1a to 1x have to do with consolidating information on System 1 classification categories and effect size ranges. Questions will be based on the similarities among study comparisons coded in Part 4 in the System 1 classifications they were coded under, and the effect size range conventions they were classified under in Part 4 (as classified under Cohen, 1988). Based on these similarities, you are to answer Yes/No questions on whether CBT changes child anxiety problems. The questions addressed in this section of Yes/No consist of four types:

> **(1)** Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?
>
> **(2)** Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?
>
> **(3)** Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?
>
> **(4)** Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

NOTE. VERY IMPORTANT, IN QUESTIONS 1-4 ABOVE, BY "SIMILAR EFFECT SIZE RANGES," WE MEAN THAT THE RANGES OF THE STUDIES ARE CLASSIFIED IN THE SAME EFFECT SIZE RANGE CATEGORIES, ACCORDING TO COHEN (1988) EFFECT SIZE CONVENTIONS, AND THE RANGE PERMUTATIONS NOTED PREVIOUSLY IN THIS MANUAL. ALSO, BY "BUT DID NOT HAVE TO," WE MEAN THAT THE RANGES EITHER COULD OR COULD NOT BE THE SAME.

Again, the coding of studies in this section consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous notes on manuals in bold in Part 4.

(1a) *Yes/No Question # 1, Applied to Total Studies.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(1b) *Yes/No Question # 2, Applied to Total Studies.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1c) *Yes/No Question # 3, Applied to Total Studies.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1d) *Yes/No Question # 4, Applied to Total Studies.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1e) *Yes/No Question # 1, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1f) *Yes/No Question # 2, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1g) *Yes/No Question # 3, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1h) *Yes/No Question # 4, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1i) *Yes/No Question # 1, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1j) *Yes/No Question # 2, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1k) *Yes/No Question # 3, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1l) *Yes/No Question # 4, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1m) *Yes/No Question # 1, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1n) *Yes/No Question # 2, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1o) *Yes/No Question # 3, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1p) *Yes/No Question # 4, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1q) *Yes/No Question # 1, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1r) *Yes/No Question # 2, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1s) *Yes/No Question # 3, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1t) *Yes/No Question # 4, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1u) *Yes/No Question # 1, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1v) *Yes/No Question # 2, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(1w) *Yes/No Question # 3, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(1x) *Yes/No Question # 4, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2) *Classifications of Treatments Using System 1, BPT for Child Conduct Problems (Item 2a through Item 2x).* In this first section of Yes/No questions for BPT for child conduct problems, we will answer Yes/No questions, based on codes we made in Part 4, under System 1. All of the questions from Item 2a to 2x have to do with consolidating information on System 1 classification categories and effect size ranges. Questions will be based on the similarities among study comparisons coded in Part 4 in the System 1 classifications they were coded under, and the effect size range conventions they were classified under in Part 4 (as classified under Cohen, 1988). Based on these similarities, you are to answer Yes/No questions on whether BPT changes child conduct problems. The questions addressed in this section of Yes/No consist of four types:

> **(1)** Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?
>
> **(2)** Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(3) Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(4) Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

NOTE. VERY IMPORTANT, IN QUESTIONS 1-4 ABOVE, BY “SIMILAR EFFECT SIZE RANGES,” WE MEAN THAT THE RANGES OF THE STUDIES ARE CLASSIFIED IN THE SAME EFFECT SIZE RANGE CATEGORIES, ACCORDING TO COHEN (1988) EFFECT SIZE CONVENTIONS, AND THE RANGE PERMUTATIONS NOTED PREVIOUSLY IN THIS MANUAL. ALSO, BY “BUT DID NOT HAVE TO,” WE MEAN THAT THE RANGES EITHER COULD OR COULD NOT BE THE SAME.

Again, the coding of studies in this section consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous notes on manuals in bold in Part 4.

(2a) *Yes/No Question # 1, Applied to Total Studies.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(2b) *Yes/No Question # 2, Applied to Total Studies.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2c) *Yes/No Question # 3, Applied to Total Studies.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(2d) *Yes/No Question # 4, Applied to Total Studies.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2e) *Yes/No Question # 1, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(2f) *Yes/No Question # 2, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2g) *Yes/No Question # 3, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(2h) *Yes/No Question # 4, Applied to Studies Using Same Manual # 1.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** “Limited Evidence for Change” or “No Evidence for Change,” **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2i) *Yes/No Question # 1, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either “Best Evidence for Change” or “Probable Evidence for Change,” **AND** provided similar effect size ranges (Yes/No)?

(2j) *Yes/No Question # 2, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2k) *Yes/No Question # 3, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2l) *Yes/No Question # 4, Applied to Studies Using Same Manual # 2.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2m) *Yes/No Question # 1, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2n) *Yes/No Question # 2, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2o) *Yes/No Question # 3, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2p) *Yes/No Question # 4, Applied to Studies Using Same Manual # 3.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2q) *Yes/No Question # 1, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2r) *Yes/No Question # 2, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2s) *Yes/No Question # 3, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2t) *Yes/No Question # 4, Applied to Studies Using Same Manual # 4.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2u) *Yes/No Question # 1, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2v) *Yes/No Question # 2, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under the System 1 classification categories of either "Best Evidence for Change" or "Probable Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(2w) *Yes/No Question # 3, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **AND** provided similar effect size ranges (Yes/No)?

(2x) *Yes/No Question # 4, Applied to Studies Using Same Manual # 5.* Are there at least two study comparisons that can be classified under **ANY** System 1 classification category that was **NOT** "Limited Evidence for Change" or "No Evidence for Change," **BUT DID NOT HAVE TO** provide similar effect size ranges (Yes/No)?

(3) *Classifications of Treatments Using System 2, CBT for Child Anxiety Problems (Item 3a through Item 3l).* In the first section of CBT for child anxiety problems, we answered Yes/No questions based on codes made in Part 4, under System 1. In this second section of Yes/No questions for CBT for child anxiety problems, we will answer Yes/No questions, based on codes we made in Part 4, under System 2. All of the questions from Item 3a to 3l have to do with consolidating information on System 2 classification categories. Questions will be based on the similarities among study comparisons coded in Part 4 in the System 2 classifications they were coded under. Based on these similarities, you are to answer Yes/No questions on whether CBT for child anxiety problems meets criteria for either System 2, Table 1 (Well-Established Psychosocial Intervention), or System 2, Table 2 (Probably Efficacious Psychosocial Intervention). The questions addressed in this section of Yes/No consist of two types:

(1) Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(2) Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

Again, the coding of studies in this section consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous notes on manuals in bold in Part 4. To help in this section, we again present both System 2 category criteria below:

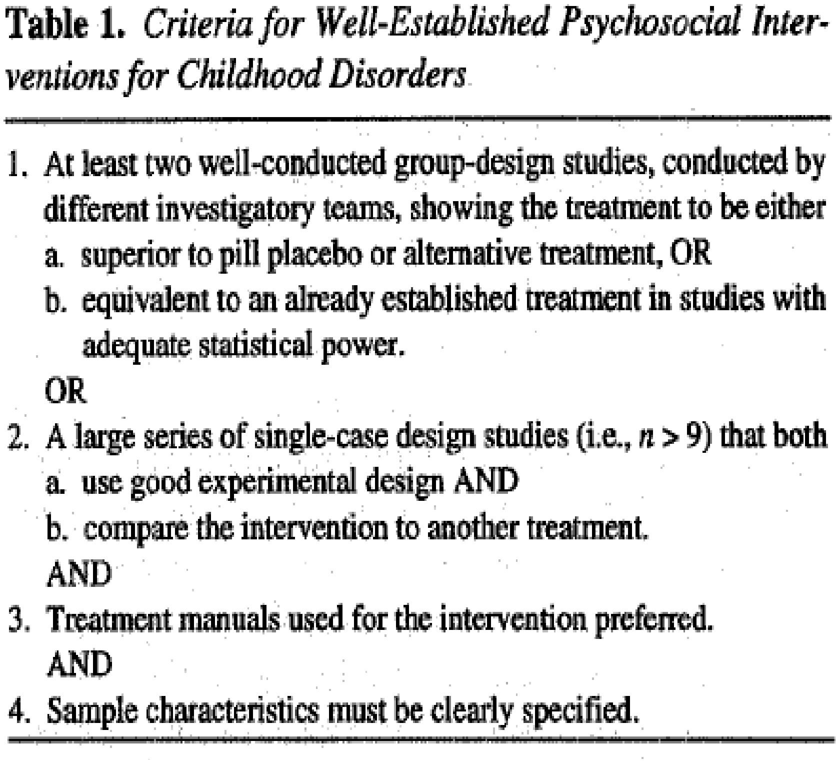


Table 1. *Criteria for Well-Established Psychosocial Interventions for Childhood Disorders*

1. At least two well-conducted group-design studies, conducted by different investigatory teams, showing the treatment to be either
   a. superior to pill placebo or alternative treatment, OR
   b. equivalent to an already established treatment in studies with adequate statistical power.

   OR
2. A large series of single-case design studies (i.e., $n > 9$) that both
   a. use good experimental design AND
   b. compare the intervention to another treatment.

   AND
3. Treatment manuals used for the intervention preferred.

   AND
4. Sample characteristics must be clearly specified.

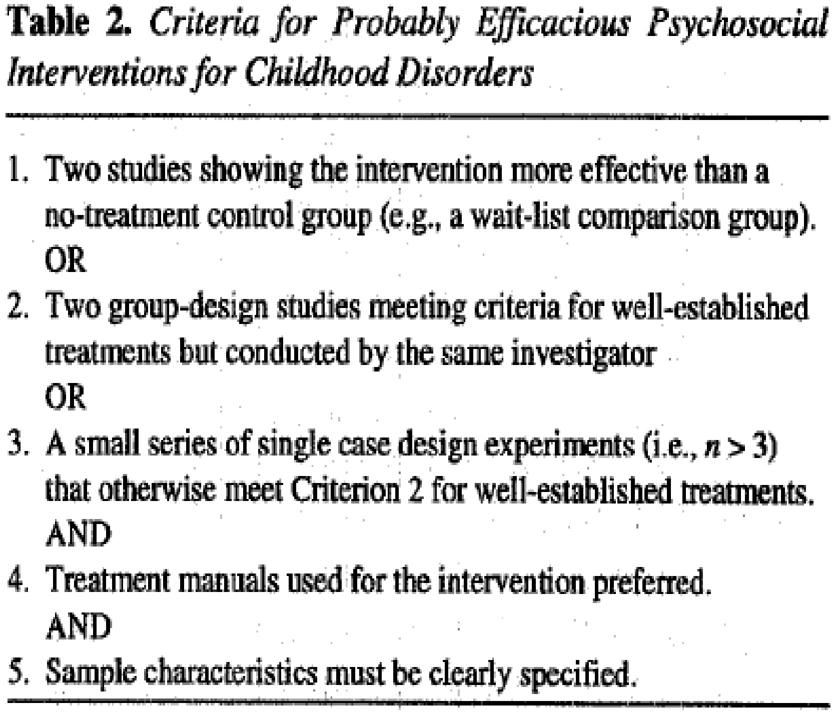


Table 2. *Criteria for Probably Efficacious Psychosocial Interventions for Childhood Disorders*

1. Two studies showing the intervention more effective than a no-treatment control group (e.g., a wait-list comparison group).
   OR
2. Two group-design studies meeting criteria for well-established treatments but conducted by the same investigator
   OR
3. A small series of single case design experiments (i.e., $n > 3$) that otherwise meet Criterion 2 for well-established treatments.
   AND
4. Treatment manuals used for the intervention preferred.
   AND
5. Sample characteristics must be clearly specified.

(3a) *Yes/No Question # 1, Applied to Total Studies.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3b) *Yes/No Question # 2, Applied to Total Studies.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(3c) *Yes/No Question # 1, Applied to Studies Using Same Manual # 1.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3d) *Yes/No Question # 2, Applied to Studies Using Same Manual # 1.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(3e) *Yes/No Question # 1, Applied to Studies Using Same Manual # 2.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3f) *Yes/No Question # 2, Applied to Studies Using Same Manual # 2.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(3g) *Yes/No Question # 1, Applied to Studies Using Same Manual # 3.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3h) *Yes/No Question # 2, Applied to Studies Using Same Manual # 3.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(3i) *Yes/No Question # 1, Applied to Studies Using Same Manual # 4.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3j) *Yes/No Question # 2, Applied to Studies Using Same Manual # 4.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(3k) *Yes/No Question # 1, Applied to Studies Using Same Manual # 5.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(3l) *Yes/No Question # 2, Applied to Studies Using Same Manual # 5.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4) *Classifications of Treatments Using System 2, BPT for Child Conduct Problems (Item 4a through Item 4l).* In the first section of BPT for child conduct problems, we answered Yes/No questions based on codes made in Part 4, under System 1. In this second section of Yes/No questions for BPT for child conduct problems, we will answer Yes/No questions, based on codes we made in Part 4, under System 2. All of the questions from Item 4a to 4l have to do with consolidating information on System 2 classification categories. Questions will be based on the similarities among study comparisons coded in Part 4 in the System 2 classifications they were coded under. Based on these similarities, you are to answer Yes/No questions on whether BPT for child conduct problems meets criteria for either System 2, Table 1 (Well-Established Psychosocial Intervention), or System 2, Table 2 (Probably Efficacious Psychosocial Intervention). The questions addressed in this section of Yes/No consist of two types:

> **(1)** Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?
>
> **(2)** Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

Again, the coding of studies in this section consists of two types: **(a)** coding similar studies within the total amount of treatment-control study comparisons; **(b)** coding similar studies that examined the outcomes of a treatment administered using the same manual or treatment protocol. Please consult the previous notes on manuals in bold in Part 4. To help in this section, we again present both System 2 category criteria below:

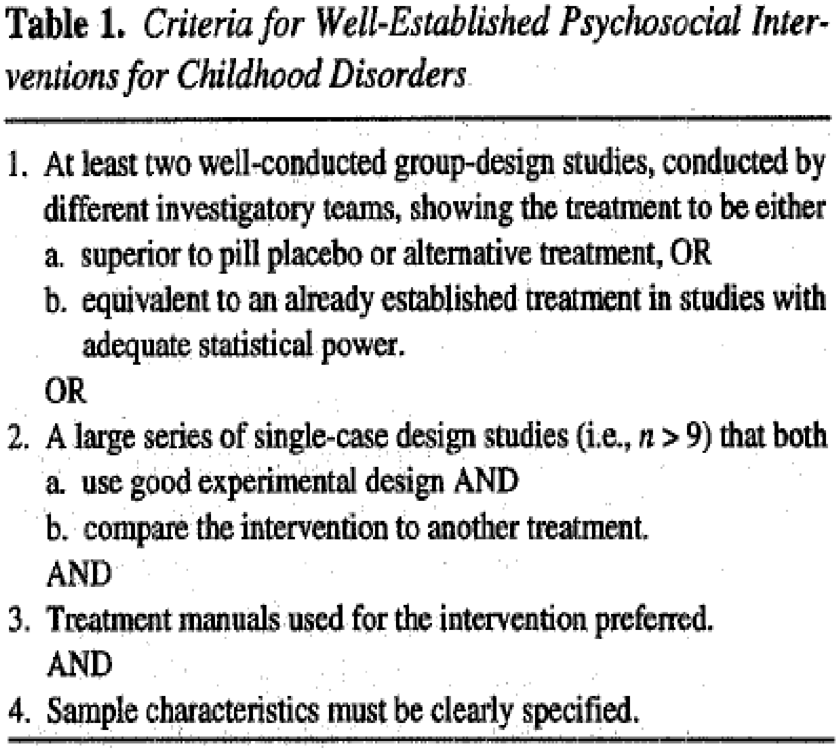


Table 1. *Criteria for Well-Established Psychosocial Interventions for Childhood Disorders*

1. At least two well-conducted group-design studies, conducted by different investigatory teams, showing the treatment to be either
   a. superior to pill placebo or alternative treatment, OR
   b. equivalent to an already established treatment in studies with adequate statistical power.

   OR
2. A large series of single-case design studies (i.e., $n > 9$) that both
   a. use good experimental design AND
   b. compare the intervention to another treatment.

   AND
3. Treatment manuals used for the intervention preferred.

   AND
4. Sample characteristics must be clearly specified.

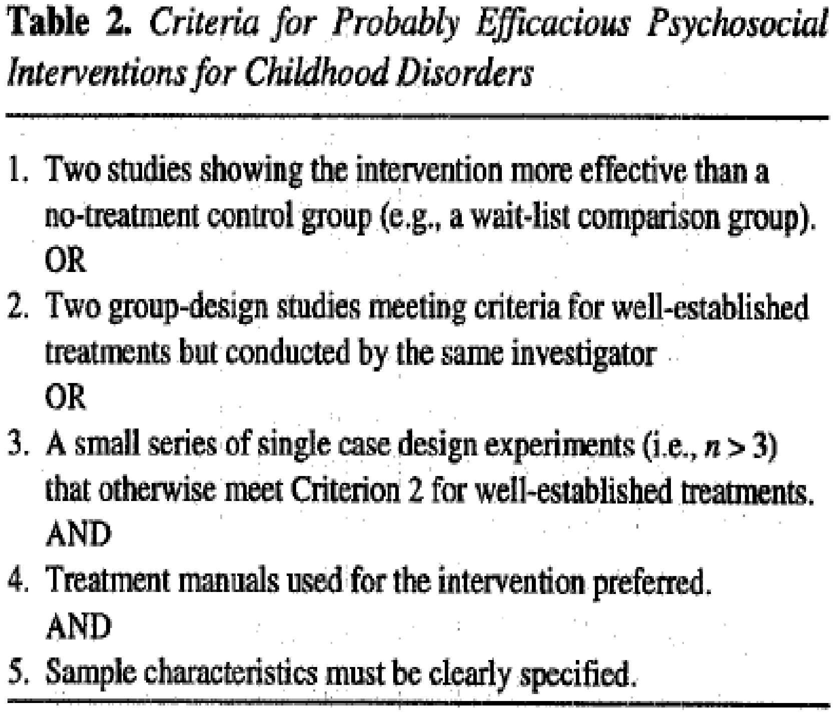


Table 2. *Criteria for Probably Efficacious Psychosocial Interventions for Childhood Disorders*

1. Two studies showing the intervention more effective than a no-treatment control group (e.g., a wait-list comparison group).
   OR
2. Two group-design studies meeting criteria for well-established treatments but conducted by the same investigator
   OR
3. A small series of single case design experiments (i.e., $n > 3$) that otherwise meet Criterion 2 for well-established treatments.
   AND
4. Treatment manuals used for the intervention preferred.
   AND
5. Sample characteristics must be clearly specified.

(4a) *Yes/No Question # 1, Applied to Total Studies.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4b) *Yes/No Question # 2, Applied to Total Studies.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4c) *Yes/No Question # 1, Applied to Studies Using Same Manual # 1.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4d) *Yes/No Question # 2, Applied to Studies Using Same Manual # 1.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4e) *Yes/No Question # 1, Applied to Studies Using Same Manual # 2.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4f) *Yes/No Question # 2, Applied to Studies Using Same Manual # 2.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4g) *Yes/No Question # 1, Applied to Studies Using Same Manual # 3.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4h) *Yes/No Question # 2, Applied to Studies Using Same Manual # 3.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4i) *Yes/No Question # 1, Applied to Studies Using Same Manual # 4.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4j) *Yes/No Question # 2, Applied to Studies Using Same Manual # 4.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

(4k) *Yes/No Question # 1, Applied to Studies Using Same Manual # 5.* Does this intervention meet criteria for a Well-Established Psychosocial Intervention (Yes/No)?

(4l) *Yes/No Question # 2, Applied to Studies Using Same Manual # 5.* Does this intervention meet criteria for a Probably Efficacious Psychosocial Intervention (Yes/No)?

APPENDIX-1

Sheet for Number of Studies

Year you are searching (please circle one): **2003** 2004 2005 2006

PsychInfo Lists:

Term Box 1: 24229

Term Box 2: 6331

Term Box 3: 23286

Total List (After combining term boxes): 902 **(DONE 1/28/06)**

MEDLINE via PubMed Lists:

Total List: 677 **(DONE 2/1/06)**

Year you are searching (please circle one): 2003 **2004** 2005 2006

PsychInfo Lists:

Term Box 1: 28661

Term Box 2: 7535

Term Box 3: 26370

Total List (After combining term boxes): 1094 **(DONE 1/29/06)**

MEDLINE via PubMed Lists:

Total List: 763 **(DONE 2/3/06)**

Year you are searching (please circle one): 2003 2004 **2005** 2006

PsychInfo Lists:

Term Box 1: 32092

Term Box 2: 9012

Term Box 3: 28928

Total List (After combining term boxes): 1382 **(DONE 9/25/06)**

MEDLINE via PubMed Lists:

Total List: 880 **(DONE 9/25/06)**

Year you are searching (please circle one): 2003 2004 2005 **2006**

PsychInfo Lists:

Term Box 1: 35596

Term Box 2: 11061

Term Box 3: 28849

Total List (After combining term boxes): 1337 **(DONE 3/15/07)**

MEDLINE via PubMed Lists:

Total List: 792 **(DONE 3/15/07)**

APPENDIX-2

Operational Definition of Cognitive-Behavioral Treatment for Childhood Anxiety Problems

(From Weisz et al., 2004)

Youth-focused cognitive behavior therapy

As Table 1 indicates, cognitive behavior therapy (CBT) has been used extensively to address youth anxiety. As the name implies, CBT entails efforts to identify and alter cognitions that contribute to the anxiety and to identify and alter maladaptive behavior (such as avoidance of feared situations) that may serve to sustain the condition. Some forms of CBT have been used in treatment focused on individual youths; other forms have been used in treatment focused on youths and their family members in various combinations (see later text). Applications of CBT at the individual-youth level have ranged from procedurally simple approaches using self-talk (typically addressing specific fears) to more complex multisession programs typically used to address multisymptom anxiety disorders.

APPENDIX-3

Operational Definition of Behavioral Parent Training for Childhood Conduct Problems

(From Weisz et al., 2004)

Behavioral parent training

As Table 4 shows, parent-focused intervention is the most extensively tested and supported form of treatment for youth conduct problems and disorders, and nearly all of the studies involved have tested behavioral parent training. Thus, significant space is devoted to discussing this approach. As the table also shows, a substantial number of the treatments that have been tested were inspired, in full or in part, by intervention procedures developed by Patterson and colleagues at the Oregon Social Learning Center. We illustrate behavioral parent training by describing some of the main elements of the Patterson approach and another method in which video vignettes are used to stimulate discussion and learning by parents who meet in groups.

The Patterson approach: Parent Management Training-Oregon

Parent Management Training-Oregon (PMTO) was developed by Patterson and colleagues through a series of pilot and research efforts in the late 1960s and early 1970s with parents of children aged 3 to 13 years. PMTO involves at least three components: **(1)** parents learn basic behavioral principles relevant to child rearing; **(2)** parents learn how to define, track, and record rates of the antisocial and prosocial behaviors they want to target; **(3)** parents are helped to design, role play, carry out, and refine behavior modification programs while continuing to record rates of target behavior to assess intervention effects. The procedures originally were used with individual parents and couples, one at a time, but to increase efficiency, some group-administered procedures also have been developed….

APPENDIX-4

List of outcome measures and acceptable subscales on measures that specifically assess childhood anxiety (Exhaustive)

Anxiety Differential: **Low scores good, High bad**

Anxiety Disorder Interview Schedule for Children (ADIS):

- Met Current Criteria for Their Primary Anxiety Disorder

-Met Current Criteria for an Anxiety Disorder

-Presence of a Posttraumatic Stress Disorder (PTSD) Diagnosis

For All: Absence good, Presence bad

Anxiety Disorder Interview Schedule for Children (ADIS):

-PTSD Symptoms (Total and All Subscales)

-Clinician Severity Rating (CSR)

-Social Phobia Severity Rating

For All: Low scores good, High bad

Behavioral Assessments (Fire Safety, number of correct responses): ***Low scores bad, High good***

Behavioral Observations (TBO; including but not limited to the following: gratuitous body movements, trembling voice, avoiding task, all subscales appropriate): **For All Subscales and Total Scores: Low scores good, High bad**

Child Behavior Checklist (CBCL): Posttraumatic Stress Disorder (PTSD) Scale: **Low scores good, High bad**

Difficult Anagrams: ***Low scores bad, High good***

Difficult Paired Associates: ***Low scores bad, High good***

Dimensions of Anxiety Index (DAI): **Low scores good, High bad**

Family Enhancement of Avoidant Responses (FEAR): **Low scores good, High bad**

Fear Inventory for Fire Safety (FIFS): **Low scores good, High bad**

Fear Survey Schedule for Children-Revised (FSSC-R): **Low scores good, High bad**

-Fear of Danger and Death Factor: **Low scores good, High bad**

-Fire Item: **Low scores good, High bad**

Fear Survey Schedule for Children-II (FSSC-II): **Low scores good, High bad**

Fear Thermometer: **Low scores good, High bad**

Fear Thermometer for Sexually Abused Children: **Low scores good, High bad**

Hindi Test Anxiety Inventory (TAI-H): **Low scores good, High bad**

Multidimensional Anxiety Scale for Children (MASC): **Low scores good, High bad**

Pictorial Dental Anxiety Scale (PDAS): **Low scores good, High bad**

Present Affect Reactions Questionnaire (PARQ): **Low scores good, High bad**

Posttask Questionnaire (PTQ):

-Anxiety Interference: **Low scores good, High bad**

-Task-Generated Interference: **Low scores good, High bad**

-Estimated Percent Time on Task: ***Low scores bad, High good***

Pulse rate: **Low scores good, High bad**

Raven’s Standard Progressive Matrices (Raven’s): ***Low scores bad, High good***

Revised Children’s Manifest Anxiety Scales (RCMAS): **Low scores good, High bad**

School attendance (if primary problem is School Refusal, a form of anxiety): ***Low scores bad, High good***

Screen for Child Anxiety Related Emotional Disorders (SCARED): **Low scores good, High bad**

Social Anxiety Scale for Children-Revised (SASC-R): **Low scores good, High bad**

Social Phobia and Anxiety Inventory for Children (SPAI-C): **Low scores good, High bad**

Social Worries Questionnaire-Pupil (SWQ-PU): **Low scores good, High bad**

Spence Children’s Anxiety Scale (SCAS)

-Total Score and Social Phobia Subscale: **Low scores good, High bad**

State-Trait Anxiety Inventory-State form (STAI-S): **Low scores good, High bad**

State-Trait Anxiety Inventory-Trait form (STAI-T): **Low scores good, High bad**

State-Trait Anxiety Inventory for Children (STAIC): **Low scores good, High bad**

Threat interpretation and response plans to ambiguity: **Low scores good, High bad**

APPENDIX-5

List of outcome measures and acceptable subscales on measures that specifically assess childhood conduct-related problems (Exhaustive)

Child Conduct Problems at Home/Mother Composite Score: **Low scores <u>good</u>, High <u>bad</u>**
Child Conduct Problems at Home/Father Composite Score: **Low scores <u>good</u>, High <u>bad</u>**
Child Conduct Problems at School Composite Score: **Low scores <u>good</u>, High <u>bad</u>**
Dyadic Parent-Child Coding System (DPICS)

- -Child Total Deviance with Mothers
- -Child Total Deviance with Fathers
- -Child Total Noncompliance

For All: **Low scores <u>good</u>, High <u>bad</u>**

Eyberg Child Behavior Inventory (ECBI):

- -Intensity Score
- -Total Problem Score

For Both: Low scores <u>good</u>, High <u>bad</u>

Parent Daily Telephone Reports or Parent Daily Reports (PDR)

- -Target Negative Behaviors
- -Number Negative Behaviors/24 Hrs.
- -Low Rate Events

For All: **Low scores <u>good</u>, High <u>bad</u>**

Peer Problem-Solving-Interaction Communication-Affect Rating Coding System (PPS-I CARE):

- -Total Negative Conflict Management: **Low scores <u>good</u>, High <u>bad</u>**

Strength and Difficulty Scale (SDQ)

- -Conduct Problems Subscale: **Low scores <u>good</u>, High <u>bad</u>**

APPENDIX-6

Examples of Methods of Analysis and Types of Statistical Tests Employed in Treatment Outcome Studies

Types of Statistical Tests

(1) *t* test: These tests may be employed to examine mean differences between conditions. However, they may also be employed to examine clinically significant change.

(2) chi square: These tests are often employed to examine diagnostic status and clinically significant change. The tests are conducted using frequencies of participants in each condition that do or do not experience diagnostic recovery/clinically significant change.

Examples of Sufficient Information to Calculate Effect Sizes and Calculate Statistical Significance Between Conditions:

(1) "Means and standard deviations are presented in Table 1. For the RCMAS, there was a significant group X time interaction (*F*[1, 86] = 5.67, *p* < .05), and follow-up tests indicated significant improvements from pre-to-posttreatment for the treatment condition, but not the control condition."
Note. The above sentence suggests all information is available in the article. The means and standard deviations are in Table 1; enough to calculate effect sizes and calculate statistical significance.

(2) "Diagnostic recovery rates were 81.1% (30/37) in the CBT condition, 75% (16/20) in the CBT + BT condition, and 48.1% (13/27) in the control condition. A significant benefit of treatment was revealed when youths treated with CBT and CBT + BT were compared to controls, (chi square [1, N = 84] = 9.29, *p* < .01)."
Note. Although the only statistical calculations in the sentence combined the two treatments to compare them to the control condition, all information is available in the sentence to compute effect sizes and calculate statistical significance. First, the frequencies of diagnostic recovery for all of the conditions are in the sentence, this is enough to compute effect sizes for these comparisons (please consult Appendix-7). Second, in Appendix-7, there are instructions to calculate statistical significance when only frequencies are presented in the article, using chi square.

(3) "Clinically significant change was measured using norms on the RCMAS, and comparing treated and control youth scores to non-clinic youths. These comparisons revealed that a significantly greater number of treated youths, as compared with control youths, showed clinically significant change, *t*(56) = 3.75, *p* < .05."
Note. In the above results sentence, no frequencies of youths exhibiting clinically significant change are provided, and the only information provided to infer statistical significance and compute effect sizes are the *t* statistic and the *p*-value. However, the *t* test and *p*-value results are enough information to infer statistical significance between conditions, and the *t* test information (*t* statistic, degrees of freedom) is enough to compute an effect size (please consult Appendix-7).

Example of Insufficient Information to Calculate Effect Sizes and Calculate Statistical Significance Between Conditions:

(1) "Diagnostic recovery rates were 81.1% in the CBT condition, 75% in the CBT + BT condition, and 48.1% in the control condition. A significant benefit of treatment was revealed when youths treated with CBT and CBT + BT were compared to controls, (chi square [1, N = 84] = 9.29, *p* < .01)."
Note. In the above results sentence, no frequencies of youths exhibiting diagnostic recovery are provided (only percentages), and the only statistical calculation made compared all treated youths to the control youths. Thus, there is insufficient information to calculate effect sizes, because often

percentages of youths cannot be converted to frequencies due to sample attrition. Also, there is insufficient information to calculate statistical significance, because the only statistical information provided compared treated youths in both treatment conditions combined to the control youths, and again, frequencies of participants in each condition were not reported to calculate chi square.

APPENDIX-7

Methods of Calculating Effect Sizes and Statistical Significance Employed in this Meta-Analysis

Calculating Effect sizes

Tests of Mean Differences Between Conditions (Means and Standard Deviations Provided in the Article):

Step 1:

$$\text{Unadjusted\_Glass's\_}\Delta = \frac{M_1 - M_2}{S_{control\_group}}$$

Where M_1 and M_2 refer to post-treatment means of treatment group and control group, respectively, and $S_{control\ group}$ refers to the post-treatment standard deviation of the control group used in the comparison.

Note. Before moving on to Step 2, please make sure that the sign of the effect size (positive or negative), relates to whether treatment condition did better than the control condition (Positive), or whether the control condition did better than the treatment condition (Negative). Please consult Appendices 4 and 5 for information on what high and low scores mean on the measure for which effect sizes comparing treatment and control conditions are being computed.

Step 2:

After the Glass's Δ effect size is calculated, this effect size must be converted to an effect size that accounts for small sample bias:

$$\text{Adjusted Glass's } \Delta = \text{ Unadjusted\_Glass's\_}\Delta * [1 - \frac{3}{(4 * \text{N}) - 9}]$$

where N is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT BOTH UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.**

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Tests of Mean Differences Between Conditions (Means and Standard Deviations Not Provided in the Article, But Statistical Results are Provided):

Step 1 (When *t* is provided):

$$\mathrm{r} = \sqrt{\frac{t^2}{t^2 + df}}$$

where t^2 is the square of the t score provided in the article, and *df* is the degrees of freedom of the *t* test. **For this formula, *df* = (n_1 – 1 + n_2 – 1).**

Step 1 (When *F* is provided):

$$\mathrm{r} = \sqrt{\frac{F}{F + df_{error}}}$$

where *F* is the *F* statistic provided in the article, and df_{error} is the degrees of freedom of the *F* test. **For this formula, df_{error} is also df_{within} and = (n_1 – 1 + n_2 – 1).**

Step 2:

VERY IMPORTANT: PLEASE NOTE THAT IN THE ABOVE CASE, THE RESULTING EFFECT SIZE MUST BE CONVERTED TO A *d*-METRIC EFFECT SIZE MEASURE (SEE LAST SECTION OF THIS APPENDIX BELOW).

Step 3:

AFTER THE CONVERSION EFFECT SIZE IS CALCULATED, THIS EFFECT SIZE MUST BE CONVERTED TO AN EFFECT SIZE THAT ACCOUNTS FOR SMALL SAMPLE BIAS:

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4 * \mathrm{N}) - 9}]$$

where *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN**

WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Tests of Diagnostic Status Between Conditions (Frequencies of Participants With and Without Disorder Provided in the Article):

Step 1:

Make a contingency table of diagnostic status between treatment and control conditions.

Diagnostic Status Table		
	Diagnosis not present	Diagnosis present
Treatment	**A (# Participants)**	**B (# Participants)**
Control	**C (# Participants)**	**D (# Participants)**

Step 2:

Once the frequencies of participants are recorded in each cell of the table, please plug the values into the following formula, which should give you Φ.

$$\Phi = \frac{(\mathrm{BC} - \mathrm{AD})}{\sqrt{(\mathrm{A}+\mathrm{B})*(\mathrm{C}+\mathrm{D})*(\mathrm{A}+\mathrm{C})*(\mathrm{B}+\mathrm{D})}}$$

VERY IMPORTANT: PLEASE MAKE SURE THAT THE SIGN OF THE RESULTING EFFECT SIZE IS POSITIVE IN CASES IN WHICH TREATMENT HAD LESS PARTICIPANTS WITH DIAGNOSIS PRESENT POST-TREATMENT, RELATIVE TO CONTROLS, AND NEGATIVE IN CASES IN WHICH TREATMENT HAD MORE PARTICIPANTS WITH DIAGNOSIS PRESENT POST-TREATMENT, RELATIVE TO CONTROLS.

Step 3:

Once Φ is calculated, it must be converted to a *d*-metric effect size measure. Please see the last section of this Appendix below for the conversion formula.

Step 4:

Once Φ is converted to *d*, the *d*-metric measure must be adjusted to correct for small sample bias.

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4*\mathrm{N}) - 9}]$$

where N is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING**

THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL “N” IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR “N” ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Tests of Diagnostic Status Between Conditions (Frequencies of Participants With and Without Disorder Not Provided in the Article, But Statistical Results are Provided):

Step 1 (When *t* is provided):

$$\mathrm{r} = \sqrt{\frac{t^2}{t^2 + df}}$$

where t^2 is the square of the t score provided in the article, and *df* is the degrees of freedom of the *t* test. **For this formula, *df* = (n_1 – 1 + n_2 – 1).**

Step 1 (When *F* is provided):

$$\mathrm{r} = \sqrt{\frac{F}{F + df_{error}}}$$

where *F* is the *F* statistic provided in the article, and df_{error} is the degrees of freedom of the *F* test. **For this formula, df_{error} is also df_{within} and = (n_1 – 1 + n_2 – 1).**

Step 1 (When chi square is provided):

$$\mathrm{r} = \sqrt{\frac{\chi^2(1)}{N}}$$

where X^2 is the chi square statistic provided in the article, and *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

Step 2:

VERY IMPORTANT: PLEASE NOTE THAT IN THE ABOVE CASE, THE RESULTING EFFECT SIZE MUST BE CONVERTED TO A *d*-METRIC EFFECT SIZE MEASURE (SEE LAST SECTION OF THIS APPENDIX BELOW).

Step 3:

AFTER THE CONVERSION EFFECT SIZE IS CALCULATED, THIS EFFECT SIZE MUST BE CONVERTED TO AN EFFECT SIZE THAT ACCOUNTS FOR SMALL SAMPLE BIAS:

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4 * \mathrm{N}) - 9}]$$

where *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.**

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Tests of Clinically Significant Change Between Conditions (Frequencies of Participants That Do or Do not Exhibit Change Provided in the Article):

Step 1:

Make a contingency table of clinically significant change between treatment and control conditions.

Clinically Significant Change Table		
	Change not present	Change present
Treatment	**A (# Participants)**	**B (# Participants)**
Control	**C (# Participants)**	**D (# Participants)**

Step 2:

Once the frequencies of participants are recorded in each cell of the table, please plug the values into the following formula, which should give you Φ.

$$\Phi = \frac{(BC - AD)}{\sqrt{(A+B)*(C+D)*(A+C)*(B+D)}}$$

VERY IMPORTANT: PLEASE MAKE SURE THAT THE SIGN OF THE RESULTING EFFECT SIZE IS POSITIVE IN CASES IN WHICH TREATMENT HAD MORE PARTICIPANTS WITH CLINICALLY SIGNFICANT CHANGE POST-TREATMENT, RELATIVE TO CONTROLS, AND NEGATIVE IN CASES IN WHICH TREATMENT HAD LESS PARTICIPANTS WITH CLINICALLY SIGNIFICANT CHANGE PRESENT POST-TREATMENT, RELATIVE TO CONTROLS.

Step 3:

Once Φ is calculated, it must be converted to a *d*-metric effect size measure. Please see the last section of this Appendix below for the conversion formula.

Step 4:

Once Φ is converted to *d*, the *d*-metric measure must be adjusted to correct for small sample bias.

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4*N) - 9}]$$

where N is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE “N” IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE “N” SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING “N” FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING**

THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Tests of Clinically Significant Change Between Conditions (Frequencies of Participants That Do or Do not Exhibit Change Not Provided in the Article, But Statistical Results are Provided):

Step 1 (When *t* is provided):

$$\mathrm{r} = \sqrt{\frac{t^2}{t^2 + df}}$$

where t^2 is the square of the t score provided in the article, and *df* is the degrees of freedom of the *t* test. **For this formula, *df* = (n_1 – 1 + n_2 – 1).**

Step 1 (When *F* is provided):

$$\mathrm{r} = \sqrt{\frac{F}{F + df_{error}}}$$

where *F* is the *F* statistic provided in the article, and df_{error} is the degrees of freedom of the *F* test. **For this formula, df_{error} is also df_{within} and = (n_1 – 1 + n_2 – 1).**

Step 1 (When chi square is provided):

$$\mathrm{r} = \sqrt{\frac{\chi^2(1)}{N}}$$

where X^2 is the chi square statistic provided in the article, and *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

Step 2:

VERY IMPORTANT: PLEASE NOTE THAT IN THE ABOVE CASE, THE RESULTING EFFECT SIZE MUST BE CONVERTED TO A *d*-METRIC EFFECT SIZE MEASURE (SEE LAST SECTION OF THIS APPENDIX BELOW).

Step 3:

AFTER THE CONVERSION EFFECT SIZE IS CALCULATED, THIS EFFECT SIZE MUST BE CONVERTED TO AN EFFECT SIZE THAT ACCOUNTS FOR SMALL SAMPLE BIAS:

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4 * \mathrm{N}) - 9}]$$

where *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

ALSO VERY IMPORTANT: **WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.**

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).

Calculating Statistical Significance

Note. Coding statistical differences between conditions involves you calculating test statistic and *p*- value information from data provided by the authors at post-treatment. Generally, tests of mean differences (where means and standard deviations of participants are provided) will be calculated using an online calculator designed to calculate *t* tests. Further, tests of diagnostic status and clinically significant change (where participant frequencies are provided of participants that either do or do not meet criteria for a diagnosis, or either do or do not experience clinically significant change) will be calculated using chi square.

VERY IMPORTANT: **Sometimes the only data available from which to calculate an effect size is a test statistic, like chi square or *t*. In those instances, the statistical test is already calculated for you. Of course, this is assuming that the statistical test that was calculated is comparing the treatment and control conditions of interest, and not some other comparison (the treatment of interest and some other treatment are combined to compare to the control condition simultaneously). If the test conducted was appropriate for calculating effect sizes for the specific treatment and control conditions being compared, please just record that statistical information, and use that information to both calculate the effect size, and determine statistical significance.**

If authors provide frequencies of participants in the groups required to calculate chi square:

Please record frequencies of participants in each group (e.g., those that did and did not recover from disorder after intervention, those that did and did not reach clinically significant change), and plug data into the chi square web calculator located here: *http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html*. A given chi square calculation will only need 2 columns, and 2 rows (a column and row for the treatment and control group being compared to each other).

If authors provide means, standard deviations, and participants in the groups required to calculate the t test:

Please record means and standard deviations and *N*s of the treatment and control conditions of participants in each group, and plug data into the *t* test web calculator located here: *http://www.graphpad.com/quickcalcs/ttest1.cfm?Format=SD*. Please make sure the calculation is set to "Enter mean, SD, and N" and "Unpaired *t* test"

VERY IMPORTANT: PLEASE MAKE SURE ANY STATISTICAL TESTS YOU CONDUCT OR CALCULATE ON THE SITES ABOVE ARE TWO-TAILED TESTS, AND NOT ONE-TAILED TESTS.

Re-Calculating Effect Sizes from an r-Family Effect Size to a d-Family Effect Size (If Needed)

Note. Re-calculating is necessary in cases in which *r* or Φ is used to calculate effect sizes, and it is required that the *r* or Φ effect size is converted to a *d*-metric, to be on the same metric as other effect sizes (Glass's∆ is a *d*-metric effect size).

Step 1:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

where *r* equals the *r*-family effect size measure (either *r* or Φ) that you wish to convert to *d*.

Step 2:

VERY IMPORTANT: AFTER THE CONVERSION EFFECT SIZE IS CALCULATED, THIS EFFECT SIZE MUST BE CONVERTED TO AN EFFECT SIZE THAT ACCOUNTS FOR SMALL SAMPLE BIAS:

$$\text{Adjusted } d = \text{Unadjusted\_d} * [1 - \frac{3}{(4 * N) - 9}]$$

where *N* is equal to the number of participants in the control group plus the number of participants in the treated group.

VERY IMPORTANT: PLEASE REPORT ALL UNADJUSTED AND ADJUSTED EFFECT SIZES ON THE CODING SHEET.

***ALSO VERY IMPORTANT:* WHEN COMPUTING THE ADJUSTED EFFECT SIZE, A CRITICAL DECISION TO BE MADE IS HOW TO DEFINE "N" IN THE EQUATION. SOMETIMES YOU CANNOT COMPUTE "N" SIMPLY BY LOOKING BACK TO THE BASELINE DATA OF HOW MANY PARTICIPANTS WERE IN THE GROUPS BEING COMPARED AND SUMMING UP THE NUMBERS FOR THE TWO GROUPS. THIS IS BECAUSE THE ANALYSES ON THE GROUPS SOMETIMES HAVE MISSING DATA, SO NOT ALL PARTICIPANTS ARE INCLUDED IN THE ANALYSES. THUS, WHEN DEFINING "N" FOR THE ADJUSTED EFFECT SIZES, PLEASE USE THE NUMBER OF PARTICIPANTS ACTUALLY PROVIDING DATA FOR THE ANALYSIS. YOU CAN DEDUCE HOW MANY PARTICIPANTS ARE IN THE ANALYSIS BY THE DEGREES OF FREEDOM (FOR *F* AND *t* TESTS, USUALLY 2 MORE PARTICIPANTS THAN THE DEGREES OF FREEDOM REPRESENTING THE WHOLE GROUP BEING ANALYZED) OR THE TOTAL "N" IN THE ANALYSIS (IN CASES OF CHI SQUARE, THIS IS USUALLY PROVIDED, EITHER IN THE ANALYSIS, OR BY SUMMING UP THE FREQUENCIES OF PARTICIPANTS BEING COMPARED). HOWEVER, IN INSTANCES IN WHICH DEGREES OF FREEDOM OR "N" ARE NOT PROVIDED, PLEASE USE GROUP PARTICIPANT DATA ACCOUNTING FOR ATTRITION (OR AS A LAST RESORT, BASELINE DATA) TO COMPUTE THE ADJUSTED EFFECT SIZE, AND PLEASE NOTE IN THE CODING SHEET THAT THIS WAS DONE TO COMPUTE THE ADJUSTED EFFECT SIZE.**

AGAIN, IF YOU ARE UNSURE ABOUT WHAT SAMPLE SIZE NUMBER TO USE, AND IF ATTRITION DATA IS ABSENT OR FOR SOME REASON CANNOT BE USED TO ESTIMATE THE POST-TREATMENT SAMPLE SIZE, PLEASE USE THE BASELINE DATA (DATA REPORTED ON SAMPLE SIZE PRE-DROPOUT), AND PLEASE NOTE ON THE CODING SHEET THAT THE BASELINE DATA WAS USED. SOMETIMES THIS BASELINE SAMPLE DATA CAN BE FOUND EVEN WHEN THE SAMPLE SIZES OF THE TREATMENT AND CONTROL GROUPS BEING COMPARED ARE NOT

REPORTED IN THE ARTICLE, SUCH AS WHEN A SINGLE TREATMENT AND A SINGLE CONTROL GROUP ARE THE ONLY GROUPS IN THE STUDY (I.E., JUST USE TOTAL BASELINE SAMPLE SIZE, USUALLY REPORTED IN ABSTRACT).