When the Evidence Says, "Yes, No, and Maybe So"

Attending to and Interpreting Inconsistent Findings Among Evidence-Based Interventions

Andres De Los Reyes and Alan E. Kazdin

Yale University

ABSTRACT—*An international, multidisciplinary effort aims to identify evidence-based treatments (EBTs) or interventions. The goal of this effort is to identify specific techniques or programs that successfully target and change specific behaviors. In clinical psychology, EBTs are identified based on the outcomes of randomized controlled trials examining whether treatments outperform control or alternative treatment conditions. Treatment outcomes are measured in multiple ways. Consistently, different ways of gauging outcomes yield inconsistent conclusions. Historically, EBT research has not accounted for these inconsistencies. In this paper we highlight the implications of inconsistencies, describe a framework for redressing inconsistent findings, and illustrate how the framework can guide future research on how to administer and combine treatments to maximize treatment effects and how to study treatments via quantitative review.*

KEYWORDS—*efficacy; effectiveness; intervention; range of possible changes; treatment*

Movements toward identifying evidence-based treatments (EBTs) or interventions encompass multiple disciplines, including dentistry, education, medicine, nursing, psychology, and social work. Scientists in each area conduct research to identify specific interventions, therapies, or programs that successfully target and change specific problem domains or behaviors (e.g., academic achievement, mood, delinquency, hypertension). Within psychology—particularly clinical, counseling, educational, and school psychology—several EBTs have been identified. Different professional groups, organizations, and task forces, as well as groups in different countries (e.g., within the European Union), states, provinces, and territories (e.g., within the United States and Canada) have developed systems delineating specific criteria for identifying EBTs. A key criterion is that the treatment outperforms a no-treatment or alternative-treatment group in randomized controlled trials. This paper elaborates on this criterion, highlights critical interpretive problems that apply to treatment research, and describes a way to redress these problems. We raise these issues within evidence-based psychotherapy specifically, but the points apply to evidence-based intervention research more generally.

Address correspondence to Andres De Los Reyes, University of Illinois at Chicago, Department of Psychiatry, Institute for Juvenile Research (MC 747), 1747 West Roosevelt Road, Room 335, Chicago, IL 60608; e-mail: areyes@psych.uic.edu.

INCONSISTENCIES IN THE EVIDENCE

Controlled trials use multiple outcome measures of a given construct and assessments of multiple constructs—sound scientific practices when defining a construct. This strategy has heightened significance for research on identifying EBTs, because a single measure rarely captures the constructs of interest: Patient outcomes and the range of domains reflecting dysfunction or well-being (e.g., positive changes in maladjustment, anxiety, impairment, mood). Thus, a single study includes multiple measures of both the same construct (e.g., depression) and related constructs (e.g., anxiety, impairment). These multiple measures vary in terms of the information source (e.g., relatives, teachers, clinicians), as well as in terms of the ways measurements are taken (e.g., symptom counts, disorder diagnoses) and examined statistically. Researchers rarely hypothesize whether some measures and not others will support the treatment. Often, it appears that researchers expect *all* measures to suggest the treatment is effective.

What if the measures do not all lead to the same conclusion? If, for example, ten measures are used, how many of these measures should support the treatment? Should two of ten

measures support it, or five of ten, or eight of ten? Currently, treatment research does not readily address these questions. This is a critical issue in EBT research, because inconsistencies often arise across assessments of adults and youths and across the many constructs treated in the clinical sciences (e.g., depression, aggression, parenting; Achenbach, 2006; De Los Reyes & Kazdin, 2005, 2006). Multiple measures are necessary and each provides reliable and valid information; it is not the case that some are "right" and others "wrong." Yet, they often lead to inconsistent conclusions.

Within studies, only some measures show that the treatment and control conditions are statistically different (e.g., De Los Reyes & Kazdin, 2006; Flannery-Schroeder & Kendall, 2000; Webster-Stratton & Hammond, 1997). Often, researchers focus on supportive measures and do not discuss the other measures or merely note that they did not "come out." Further, between two or more studies of the same treatment, measures that support and do not support the treatment in one study do not necessarily lead to the same conclusions in other studies (e.g., Barrett, Dadds, & Rapee, 1996; Kendall, 1994; Kendall et al., 1997). Therefore, at the end of controlled trials, conclusions can range from stating that the treatment is evidence-based to stating that it is not evidence-based, or to stating that the evidence is mixed and dependant on the measure relied on to define treatment outcomes (De Los Reyes & Kazdin, 2006).

There has been insufficient recognition of inconsistent evidence, and no model exists to integrate inconsistencies that accounts for all of the evidence. It is possible to acknowledge inconsistencies and still use the evidence to identify EBTs. Indeed, inconsistent findings might signify important circumstances in which evidence suggests treatments are effective and circumstances in which evidence is inconclusive. For instance, consistent findings based on informants that observe behavior in one context (e.g., a mother observing her child at home), and inconsistent findings based on other informants that observe behavior in another context (e.g., a teacher observing that same child at school) may suggest where an intervention may yield particularly robust outcomes (home-based rather than school-based behavior). One way of addressing inconsistencies in the identification of EBTs is to devise a plan for identifying patterns in evidence that reveal the ways in which treatments are most effective.

THE RANGE OF POSSIBLE CHANGES MODEL

The Range of Possible Changes (RPC) Model was designed to consider within- and between-study consistencies to identify EBTs. By "range," we mean the myriad conclusions that might be drawn from multiple findings that are discrepant in their support (or lack thereof) of a particular treatment's effects. This includes treatment literatures that often employ a single measure or source to gauge treatment effects (e.g., smoking cessation, weight loss). Indeed, in these treatment literatures, the methods by which outcomes are quantified are often arbitrary (Blanton & Jaccard, 2006), suggesting that even single outcomes can and ought to be examined in multiple ways.

The model provides a classification system that identifies EBTs based in part on whether multiple or specific outcome methods consistently yield similar conclusions. Within this system are categories that classify the many different kinds of studies that produce evidence for treatments (Table 1). Broadly, the categories span classifications of studies that find consistent evidence across multiple ways of gauging outcomes (e.g., Best Evidence for Change), consistent evidence when employing specific outcome methods (e.g., Evidence for Measure- or Method-Specific Change), and inconsistent evidence (e.g., Limited Evidence for Change; De Los Reyes & Kazdin, 2006). Further, the categories can be applied to classifying evidence, depending on what is targeted for treatment. In other words, one can classify evidence based on multiple measures that represent the same outcome domain (e.g., multiple symptom reduction measures, multiple risk factor measures). Most critically, the RPC Model can be used to examine whether two studies of the same treatment yield consistent evidence between them. An example would be two studies examining whether a particular treatment reduces symptoms of anxiety. If the studies could both be classified within the same category (e.g., Best Evidence for Change), then they may be classified as providing consistent evidence for the reduction of anxiety symptoms.

In addition, the model acknowledges that outcomes might be tested in multiple ways. Specifically, outcomes are often evaluated by examining statistical differences between treatment and control conditions, yielding a limited set of possible findings (e.g., treatment is effective, evidence is inconclusive, treatment makes people worse). Indeed, the classification categories described in Table 1 are based on this method. However, another method assesses *how much* of a difference exists between conditions (e.g., effect size, or the degree of difference between the average scores of treatment and control participants). Combining these two methods might reveal nuances in a treatment's effectiveness. For example, a study's evidence might meet criteria for the Best Evidence for Change category (Table 1) and yet have observed magnitudes of change ranging from small to large. Thus, the RPC Model addresses the issue of multiple methods of testing outcomes by incorporating treatment outcomes classifications based on categorical statistical differences with evaluations of the range of outcomes based on degree of statistical differences (for a discussion of measurement reliability and statistical power issues see De Los Reyes & Kazdin, 2006).

ADVANCES AND FUTURE DIRECTIONS

Prior research has identified EBTs and yet has not accounted for inconsistent findings. However, inconsistencies may reveal important information of treatment effects: They may highlight both the variety of ways that a treatment may change behaviors and the specific circumstances in which a treatment may be effective

TABLE 1
Description and Criteria of Range of Possible Changes (RPC) Model Categories

Category	Criteria
Best Evidence for Change	At least 80% of the findings from three or more informants, measures, and analytic methods show differences, and at least three findings were gleaned from each of the informants, measures, and methods. There is no clear informant-specific, measure-specific, or method-specific pattern of findings. The evidence suggests the intervention successfully targets the construct.
Evidence for Probable Change	More than 50% of the findings from three or more informants, measures, and analytic methods show differences, and at least three findings were gleaned from each of the informants, measures, and methods. There is no clear informant-specific, measure-specific, or method-specific pattern of findings. The evidence suggests the intervention probably changes the targeted outcome domain, yet future work ought to examine why inconsistencies occurred.
Limited Evidence for Change	Either 50% or less of the findings from three or more informants, measures, and analytic methods show differences, or less than the grand majority (less than 80%) of findings from specific informant's ratings, measures, and/or methods show differences. Any differences found are either scattered across outcomes from multiple informants, measures, or methods, or are not found predominantly on outcomes from specific informants, measures, and/or methods. The evidence is inconclusive.
No Evidence for Change	No differences are observed. The evidence is completely inconclusive.
Evidence for Informant-Specific Change	Differences are found on the grand majority (80%) of ratings provided by specific informant(s), and at least three findings were gleaned from the informant(s) for which specificity of findings were observed. The evidence suggests the treatment might change the domain when it is exhibited in specific situations or in interactions with specific informant(s).
Evidence for Measure- or Method-Specific Change	Differences are found on the grand majority (80%) of specific measure(s) or analytic method(s), and at least three findings were gleaned from the measure(s) or method(s) for which specificity of findings were observed. The evidence suggests the intervention might change the domain when it is measured with specific kinds of measure(s), method(s), or both.

Note. Adapted from De Los Reyes & Kazdin, 2006. By ''informants'' we mean reporters of outcomes (e.g., self, spouse or significant other, clinician, laboratory observer, biological, institutional records); by ''measures'' we mean ways to assess outcomes (e.g., questionnaire or symptom-count measures, laboratory observations, diagnostic interviews); by ''analytic methods'' we mean statistical strategies (e.g., tests of mean differences, tests of diagnostic status).

(Table 1). The RPC Model addresses inconsistencies and reveals directions for future research that could lead to a greater understanding of how to administer and combine treatments to maximize their effects and how to conduct meta-analytic reviews of treatment research.

First, the RPC Model identifies the circumstances in which treatments might produce consistent effects. For instance, consider a treatment that the evidence suggests produces robust effects within specific circumstances (e.g., symptom reduction, at school or with peers) and inconsistent effects within other circumstances (e.g., diagnostic remission, at home). With this evidence, researchers have an increased understanding of how to administer that treatment in future studies (e.g., where effects were consistently observed). Further, researchers have a greater understanding of how long that treatment ought to be administered (e.g., enough to produce symptom reductions, longer to produce both symptom reductions and diagnostic remission). Therefore, the RPC Model guides knowledge of treatment effects, leading to sensible decision making as to where and how to administer treatments.

Second, the RPC Model identifies two potentially fruitful methods for combining treatments. Broadly, one might conceptualize combining treatments such that each treatment produces consistent effects that the other treatment does not produce. This strategy is like fitting two puzzle pieces together, where each piece fills in the gaps left open by the other piece. Specifically, one strategy might involve combining two or more treatments that are identified as producing consistent effects in different domains of the same construct. An example might be a combined protocol including a treatment that both consistently produces effects on symptom outcome measures and inconsistently produces effects on risk-factor outcome measures with another treatment that consistently produces effects on risk-factor outcome measures and not on symptom outcome measures.

Another method might involve two or more treatments that are identified as producing consistent effects in different contexts or circumstances within the same domain (Table 1). For instance, one might combine a treatment that produces consistent symptom reductions on school-based and not home-based measures with a treatment that produces consistent symptom reductions on home-based and not school-based measures. Therefore, the RPC Model guides the development of cost-effective methods of combining treatments so that effects are not redundant between treatments in a combined protocol.

Third, the RPC Model informs future meta-analytic reviews of treatment research. Indeed, traditional meta-analytic reviews

have identified effects of specific treatment techniques by averaging effects multiple times—not only within studies but also between studies of the same or similar techniques (e.g., Matt, 1989; Stice & Shaw, 2004). However, with average treatment effects, it remains unclear whether consistent evidence is found within any one study or between any two studies. For instance, a sample of treatment studies might on average yield large treatment effects. Yet, that sample might include multiple studies that only yielded statistically significant effects on half of their outcome measures, with no two studies yielding the same ranges of magnitudes of effects (e.g., no two studies suggesting effects ranged from medium to large). Further, even procedures that statistically correct for potentially biasing factors in effect-size estimates (differences in integrity of administration of treatment, differences in reliability of measures; Hunter & Schmidt, 2004) still often apply these corrections at an aggregate level (e.g., across outcomes within a study or across averages of outcomes within an entire study sample). Aggregate measures and their corrections do not necessarily yield evidence on whether individual measures within and between studies are replicating the same effect or consistently suffering from the same biasing factors (De Los Reyes & Kazdin, 2006).

The RPC Model might be used to study evidence via meta-analysis, by employing both categorical (Table 1) and continuous (effect-size) measures of treatment effects. For example, within a sample of studies of the same treatment, one could both classify each study categorically using the RPC Model categories and calculate effect sizes for each outcome to determine the range of effects observed for each study (i.e., highest and lowest effect sizes). With this critical information, one can address a number of pertinent research questions. For example, one can examine whether multiple studies are both consistently classified in the same RPC Model category and show similar ranges of treatment effects (e.g., two or more studies classified in the Evidence for Probable Change category, exhibiting medium-to-large treatment effects). Further, one could examine moderators of both RPC Model categorical classifications and moderators of the upper and lower limit effects observed within each study. For instance, one could study whether sample (gender, age), treatment (individual vs. group), and methodological (reliability of measures) characteristics are related to the likelihood that a study would be classified in a particular RPC Model category or related to the average range or distance between the highest and lowest effect sizes observed within studies. Additionally, the framework's use of effect-size measures makes it possible to use versions of statistical correction procedures to account for differences among studies in treatments examined and differences among outcome measures in their reliability or other measurement properties (Hunter & Schmidt, 2004). Thus, one can study treatments meta-analytically and still account for important information on the consistency in treatment effects, as well as identify moderators of within- and between-study consistency.

CONCLUSIONS

The movement toward identifying EBTs advances a research literature that spans multiple disciplines and types of interventions in mental health, physical health, and education. Our aim in this article is to enhance the already remarkable gains made in EBT research and the broader EBT movement. In the practice of clinical psychology, non-EBTs for adults and youths continue to be used when EBTs that target the same behaviors are available. Although a given study might reveal inconsistent outcomes—and this raises significant issues—this ought to be presented in the context of a key reality: Hundreds of "evidence-less" treatments are being administered to patients (Kazdin, 2000), and some evidence, although inconsistent, is clearly better than none. We do not advocate non-EBTs where EBTs are available.

A critical interpretive issue requires further attention: In a given study and across studies that replicate that original study, some measures show a change and others do not. This reality applies to treatments for both adults and youths and encompasses the range of behaviors targeted in research. There has been tacit selection of the measures that show change. In part, this selection is driven by basic-science issues, in that "null and negative effects" are difficult to interpret and can arise for myriad reasons (e.g., low statistical power or small sample size, poor measure reliability). However, statistically significant and positive effects might also be difficult to interpret and can arise for multiple reasons. Null effects can be real (i.e., reflect that no change occurred), just as much as significant changes on measures could be attributable to chance fluctuations in outcomes.

The RPC Model takes into account inconsistencies, and employing the framework will allow researchers to draw reliable and valid conclusions amidst them. Further, the RPC Model yields interesting directions for future research on understanding intervention effects and how to maximize them. We encourage future research that uses the RPC Model to evaluate (a) the circumstances in which interventions produce the most consistent effects, (b) ways of combining interventions, and (c) intervention effects via meta-analytic review. More than a single model, we encourage further work on the matter of inconsistencies and how they ought to be integrated to draw conclusions from EBT research.

Recommended Reading

Achenbach, T.M., Krukowski, R.A., Dumenci, L., & Ivanova, M.Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*, 361–382. Documented the general importance of informant discrepancies to adult clinical assessments.

Achenbach, T.M., McConaughy, S.H., & Howell, C.T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psycho-*

logical Bulletin, *101*, 213–232. A seminal meta-analysis that identified informant discrepancies as a general clinical child assessment issue.

Achenbach, T.M. (2006). (See References). A brief review of the implications of informant discrepancies for clinical assessment.

De Los Reyes, A., & Kazdin, A.E. (2005). (See References). Advances a theoretical framework to explain why informant discrepancies exist in clinical child assessments.

De Los Reyes, A., & Kazdin, A.E. (2006). (See References). Discusses the RPC Model in more detail than the current article.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. This paper provides a general treatment of meta-analysis and its methodology.

Acknowledgments—This work was supported, in part, by National Institute of Mental Health (NIMH) Grant MH67540 (Andres De Los Reyes) and by NIMH Grant MH59029 (Alan E. Kazdin). We are very grateful to Shannon M.A. Kundey for her extremely helpful comments on a previous version of this paper.

REFERENCES

Achenbach, T.M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*, 94–98.

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.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27–41.

De Los Reyes, A., & Kazdin, A.E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*, 483–509.

De Los Reyes, A., & Kazdin, A.E. (2006). Conceptualizing changes in behavior in intervention research: The range of possible changes model. *Psychological Review*, *113*, 554–583.

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.

Hunter, J.E., & Schmidt, F.L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Kazdin, A.E. (2000). *Psychotherapy for children and adolescents: Directions for research and practice*. New York: Oxford University Press.

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

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

Matt, G.E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin*, *105*, 106–115.

Stice, E., & Shaw, H. (2004). Eating disorder prevention programs: A meta-analytic review. *Psychological Bulletin*, *130*, 206–227.

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.