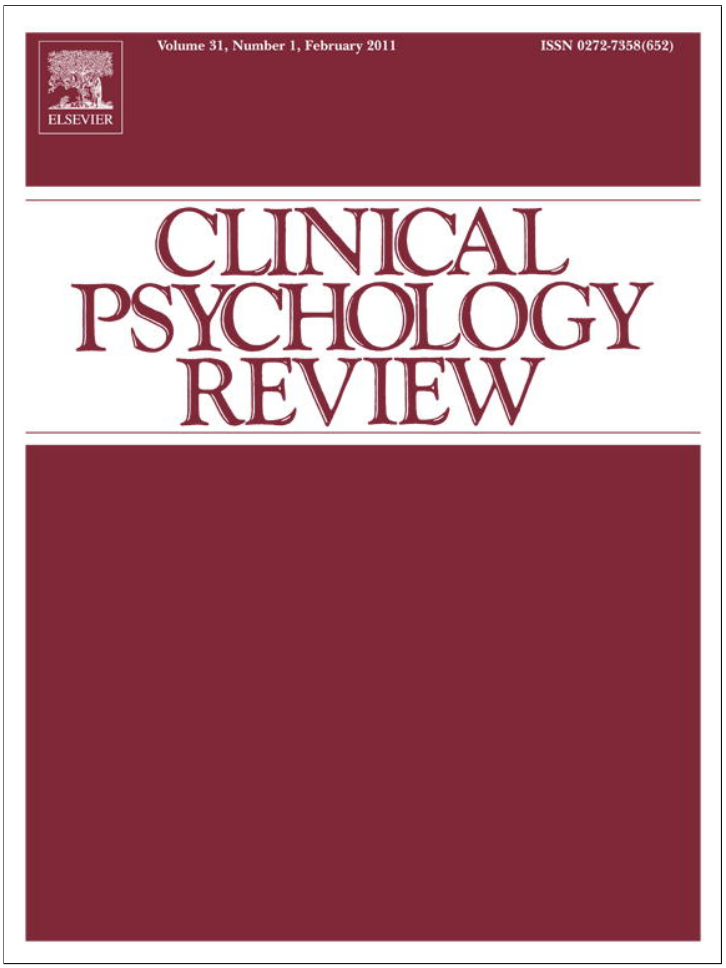

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Clinical Psychology Review 31 (2011) 829–838

Contents lists available at ScienceDirect

Clinical Psychology Review

The end of the primary outcome measure: A research agenda for constructing its replacement ☆

Andres De Los Reyes [a,*], Shannon M.A. Kundey [b], Mo Wang [a]

[a] University of Maryland at College Park, Department of Psychology, Biology–Psychology Building, College Park, MD 20742, United States
[b] Hood College, Department of Psychology, 401 Rosemont Avenue, Frederick, MD 21701, United States

ARTICLE INFO

Article history:
Received 28 November 2010
Received in revised form 25 March 2011
Accepted 27 March 2011
Available online xxxx

Keywords:
Controlled trial
Efficacy
Primary outcome measure
Standardized replication rate
Treatment

ABSTRACT

No definitive or *gold standard* outcome measure exists to test the efficacy of the mental disorder treatments examined within randomized controlled trials. As a result, researchers often evaluate efficacy via multiple outcome measures administered within a single controlled trial. This practice commonly yields inconsistent findings as to a treatment's efficacy. To address the issue of inconsistent findings, increasingly (and paradoxically) controlled trials include designations of a single measure as a *primary outcome* and other measures as *secondary outcomes*. In this paper, we review recent work highlighting the limitations of this approach to testing efficacy. In discussing how these limitations outweigh the strengths of the primary outcome method, we argue that this method needs to be replaced with an approach that addresses its limitations. In doing so, we outline the basic principles of a research agenda to develop such a replacement approach. The approach (Standardized Replication Rate [SRR] Approach) would focus on the extent to which multiple outcome measures within a controlled trial yield replicable effects, relative to the characteristics of the outcome measures and the treatment(s) examined within the trial. A research agenda focused on developing the SRR Approach would increase accountability for both reporting and interpreting controlled trials evidence.

Contents

☆ We are very grateful to Stephen Hinshaw and Bethany Teachman for their comments on previous versions of this article.

* Corresponding author at: Comprehensive Assessment and Intervention Program, Department of Psychology, University of Maryland, Biology/Psychology Building, Room 3123H, College Park, MD 20742, United States. Tel.: +1 301 405 7049; fax: +1 301 314 9566.
E-mail addresses: adelosreyes@psyc.umd.edu (A. De Los Reyes), kundey@hood.edu (S.M.A. Kundey), mwang@psyc.umd.edu (M. Wang).

0272-7358/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cpr.2011.03.011

Researchers have yet to identify definitive and cost-effective behavioral or biological markers for the specific mental disorders delineated in classification manuals (e.g., attention-deficit hyperactivity disorder [ADHD]; conduct disorder; major depression; social anxiety; and substance dependence; American Psychiatric Association, 2000). As a result, a key cornerstone of best practices in assessing and diagnosing mental disorders is the use of multiple measures of these disorders (e.g., multiple measurement methods and information sources) (Hunsley & Mash, 2007). In fact, for treatments developed for use with children and adolescents (i.e., collectively referred to as "children" unless otherwise specified) and/or adults, investigators commonly employ multiple outcome measures within controlled trials testing these treatments (see De Los Reyes & Kazdin, 2006; Weisz, Jensen Doss, & Hawley, 2005). One of the most robust observations in clinical science is that of multiple measures of the same disorder and/or its symptoms yielding different measurement outcomes (Achenbach, 2006). As a result, use of multiple outcomes within controlled trials often translates into inconsistent findings as to the efficacy of the treatment or treatments under investigation (De Los Reyes & Kazdin, 2008).

Researchers have made many attempts to develop strategies to address inconsistent findings within controlled trials, and yet no method exists that researchers agree properly addresses these inconsistencies (see De Los Reyes & Kazdin, 2005, 2006). With this in mind, one method that has gained favor in recent years is the *primary outcome method*. Specifically, within controlled trials testing treatments developed for both adults and children, researchers often select an *a priori* measure to represent overall outcomes or treatment response in a trial and deem it a *primary outcome measure* (e.g., Bowden et al., 2000; Hazell & Stuart, 2003; Papakostas, Mischoulon, Shyu, Alpert, & Fava, 2010). Researchers often include additional measures that assess the same or similar domains assessed by the primary measure; these measures are literally referred to as *secondary outcome measures* (e.g., Pettinati et al., 2010; Thurstone, Riggs, Salomonsen-Sautel, & Mikulich-Gilbertson, 2010).

How researchers designate primary and secondary outcome measures varies across literatures. For example, selection of the primary measure may occur through consensus. That is, investigators testing treatments for the same clinical condition may hold a meeting that results in an agreed-upon primary measure and criteria for gauging treatment response (e.g., adulthood major depressive disorder; see Frank et al., 1991; Zimmerman et al., 2006). Alternatively, selection of a primary measure may occur through precedent, in which investigators select a primary measure based on which measure has been most widely used in previous trials testing treatments for similar conditions (e.g., childhood anxiety disorders; see Birmaher et al., 2003; Clark et al., 2005; Wagner et al., 2004). Regardless of the method by which researchers designate primary and secondary outcome measures, both types of outcome measures often assess efficacy on identical domains (e.g., symptom presentation of the disorder targeted for treatment). Importantly, at the conclusion of the trial often all outcomes are analyzed when evaluating efficacy. Stated another way, it is a matter of convention to report the findings from analyses of both primary and secondary outcome measures, despite the fact that the main rationale for identifying a primary outcome is to rely on one metric to test efficacy.

In many respects, when using the primary outcome method, controlled trials researchers in the mental health field are following the design guidelines for trials conducted in the general medical sciences (see Boutron, Dutton, Ravaud, & Altman, 2010). Indeed, the primary outcome method has become a cornerstone of controlled trials study design. As evidence of this, consider that when the results of a controlled trial are registered within a public database – a prerequisite for publication in journal outlets edited by members of the International Committee of Medical Journal Editors – designations of the *primary* and *secondary* outcome measures must be included (De Angelis et al., 2004). Additionally, use of this methodology enjoys three key strengths. First, identifying a primary measure (theoretically) holds a researcher accountable to evaluating the treatments under investigation based on one measure, even when other outcome measures were collected and are available to test efficacy. Second, the primary outcome method greatly emphasizes parsimony and rapid communication of research findings. That is, investigators administer one primary outcome measure that yields one conclusion, greatly facilitating dissemination of findings to the public and streamlining recommendations on best practices for clinicians. Third, use of the primary outcome method eliminates the need to statistically correct for examining multiple outcomes within a controlled trial. Thus, arguably investigators have implemented the primary outcome method with the best intentions: to hold scientists to the highest standards when executing controlled trials, collecting outcomes data, making sense of these data, and communicating findings to the public.

Despite these strengths, when applied to controlled trials in the mental health field, use of the primary outcome method brings with it a variety of limitations. We argue that these limitations outweigh the strengths of using this method to assess treatment response. To be clear, we do not think that as a field we should place blame on ourselves for using this method. This is because: (a) only recently has empirical work highlighted the key limitations of the method and (b) no system currently exists to replace this method. That being said, we advance the controlled trials research literature in three ways. First, we outline, review, and illustrate the major limitations of the primary outcome method for assessing treatment response within controlled trials testing treatments for specific mental disorders. Second, in outlining these limitations we advance a research agenda for developing and implementing an eventual replacement for the primary outcome method as used in the mental health field. Third, we highlight potential challenges in revising outcome measure methodologies and outline recommendations for addressing these challenges in future controlled trials research.

1. Limitations of the primary outcome method in controlled trials testing treatments for mental disorders

Given the strengths of the primary outcome method, it is important to, first, review the limitations of the approach, and second, to discuss whether the ratio of strengths-to-limitations of the primary outcome method supports either its continued use or its replacement with an approach addressing these limitations. We have already identified two limitations that we will not review in further detail. Specifically, investigators use the primary outcome method: (a) without a definitive mechanism by which to identify a *gold standard* measure to assess efficacy; and (b) even when multiple psychometrically sound outcome measures exist to test treatment efficacy (i.e., existence of multiple viable *primary outcome measures*). These limitations are both self-evident and have been discussed elsewhere

(De Los Reyes & Kazdin, 2006). Recent research has uncovered four other important limitations of the primary outcome method.

1.1. The Consequences of Instances in Which Primary Outcome Measures Yield Null Findings

Recent work masterfully highlights perhaps the greatest limitation of the primary outcome method: When primary outcome measures yield null findings, basic human judgment processes may interfere with researchers' abilities to properly communicate these findings. Specifically, a recent meta-analysis of controlled trials published in December 2006 examined a representative sample of those studies reporting null effects of the investigated treatment based on the primary outcome measure ($n = 72$ studies of 616 published reports; Boutron et al., 2010). In their review, the authors identified studies in which investigators engaged in reporting strategies highlighting beneficial effects of the investigated treatment despite null findings based on the primary outcome measure (i.e., *spin*). The authors found that over 68% of studies used *spin* strategies in the Abstract of the article, over 60% in a section of the main text (i.e., Results, Discussion, and Conclusions), and over 40% in at least two main text sections. In the face of findings that were inconsistent with investigators' preconceived notions (i.e., hypothesis that the treatment *worked*), investigators often found ways to *make it look* as though the evidence supported the treatment's efficacy (e.g., emphasizing evidence based on secondary outcomes).

On the surface, these findings seem to indicate that current ethical standards for reporting outcome findings are not sufficiently strict, and that stricter standards would address these issues. We argue that these findings cannot be attributed to investigators' poor ethical standards. Rather, these findings reflect basic principles of human judgment known for over three decades. Specifically, when people hold preconceived notions regarding a given phenomenon (e.g., the efficacy of a treatment) and observe empirical evidence contrary to these notions (e.g., null effects on outcome measures evaluating the treatment), they become more biased in favor of these preconceived notions (Lord, Ross, & Lepper, 1979). This effect often manifests itself through the increased tendencies of people to critique the methodological rigor of the empirical evidence that is inconsistent with their preconceived notions (Lord et al., 1979).

In the case of controlled trials research, investigators often conduct trials with a preconceived notion as to the results (i.e., a research hypothesis), and the primary outcome measure serves as evidence that either confirms or fails to confirm that notion. However, as mentioned previously, controlled trials often include not only the primary outcome measure but secondary outcome measures that researchers could also use to test the treatment's efficacy. In essence, the basic science on human judgment would predict disastrous consequences from the contexts within which the primary outcome method is used in controlled trials research. That is, this set of circumstances creates the possibility that if an investigator identifies null findings on their primary measure, there are secondary measures available for the investigator to *fall back on* if the primary measure does not *work out*. Thus, investigators are at risk for revising their outcome interpretation strategies and focusing on secondary outcomes that support treatment efficacy, when their primary outcome yields evidence that is inconsistent with their expectations. In sum, although researchers choose their primary measure *a priori*, they often fail to treat the null findings related to their primary measure in an *a priori* manner. Rather, they use post hoc interpretations drawn from secondary measures to test their *a priori* hypotheses.

1.2. Primary Outcome Measures Often Rely on a Single Information Source

Another key limitation of the primary outcome method is that primary outcome measures in both the adult and child treatment literatures often rely on a single informant's report, typically a clinician (e.g., Guy, 1976; Hamilton, 1960; Scahill et al., 1997; Shear et al., 2001; Spearing, Post, Leverich, Brandt, & Nolen, 1997; Young, Biggs, Ziegler, & Meyer, 1978; Zaider, Heimberg, Fresco, Schneier, & Liebowitz, 2003). With few exceptions and in studies of treatments for both adults and children, outcome measures based on one informant's report will yield research findings that are inconsistent with findings based on other informants' reports (e.g., Casey & Berman, 1985; De Los Reyes & Kazdin, 2009; Koenig, De Los Reyes, Cicchetti, Scahill, & Klin, 2009; Lambert, Hatch, Kingston, & Edwards, 1986; Ogles, Lambert, Weight, & Payne, 1990; Weisz, McCarty, & Valeri, 2006).

Consequently, a crucial limitation of the primary outcome method is that often a researcher could defensibly implement any one of a number of information sources to complete the primary outcome measure. For instance, a plethora of outcome measures exist for use within controlled trials testing treatments for adulthood mood disorders. Specifically, these outcomes are based on information from such sources as semi-structured clinical interviews and patient self-reports; many with psychometric support for their use as outcome measures (Joiner, Walker, Pettit, Perez, & Cukrowicz, 2005). In fact, these measures have long been used as treatment outcome measures (Frank et al., 1991; Lambert et al., 1986; Zimmerman et al., 2006). Thus, the primary outcome method ignores the high likelihood that findings based on a single informant's outcome report will not replicate across other reliable and valid informants' outcome reports.

1.3. Primary outcome measures differentially relate to multiple informants' reports

It is important to note that despite the fact that clinician ratings of treatment response are typically the measure in a trial identified as the primary outcome, interpretations of the efficacy of interventions are not made specific to the characteristics of the measure (e.g., clinical interview) or the informant completing the measure (e.g., clinician). Rather, findings based on these measures are used to gauge treatment efficacy in a global sense (De Los Reyes & Kazdin, 2006). That is, for treatments for both adults and children, investigators (and treatment guidelines) largely do not qualify their interpretations of the evidence of a treatment's efficacy based on the characteristics of the supportive evidence (see American Psychological Association Interdivisional Task Force on Child & Adolescent Mental Health, 2007; Blue Cross & Blue Shield of Texas, 2007; De Los Reyes, Alfano, & Beidel, 2011; Chambless & Ollendick, 2001).

On the surface, one might surmise that clinician ratings as primary outcome measures would adequately represent *global* estimates of efficacy. Indeed, in controlled trials testing treatments for both children and adults, investigators use clinician ratings based on reports or information provided by multiple informants (e.g., spouses and patients in the case of adults; parents, children, and teachers in the case of children; Guy, 1976; Hamilton, 1960; Knopman, Knapp, Gracon, & Davis, 1994; Niederhofer, Staffen, & Mair, 2003; Scahill et al., 1997; Shear et al., 2001; Spearing et al., 1997; Young et al., 1978; Zaider et al., 2003). Thus, if clinician-specific measures are based on multiple informants' reports, then estimates of treatment response should represent these multiple perspectives. Further, presumably clinicians conducting work in a controlled trial receive training in assessing and diagnosing the condition treated in the trial. As such, their reports should converge with other reliable and valid reports from informants trained to assess the same condition, such as independent laboratory observers of the patient.

Contrary to these assumptions, however, prior work indicates that clinician reports do not reflect *global* estimates of treatment efficacy. That is, when interpreted as estimates of treatment response, clinician reports are not necessarily representative of other reports. For example, when assessing children, clinicians often believe that

multiple informants vary in whether they are capable of providing reliable and valid reports of children's behavior. In other words, depending on the problem being assessed (e.g., internalizing versus externalizing), clinicians believe that certain informants are optimal informants relative to any other available informants (see Loeber, Green, & Lahey, 1990).

Needless to say, the act of identifying certain informants as *optimal* informants lies in stark contrast with the lack of a definitive (or empirical) basis by which to identify optimal informants. Indeed, there is a wealth of evidence that the multiple informants from whom clinicians take reports of children's behavior (e.g., parents; teachers; and children) provide reliable and valid reports of children's behavior (e.g., Achenbach & Rescorla, 2001; Hunsley & Mash, 2007; Mash & Hunsley, 2005). Yet, these views might account for recent evidence indicating that when informants disagree in their reports clinicians systematically *side* with one informant's report over other reports. For instance, parents and children rarely agree on what behaviors to target in treatment, and clinicians more often agree with the parent when the targeted problem deals with the child's behavior (Hawley & Weisz, 2003). Similar findings have been observed for parent, adolescent, and clinician reports of functional impairment (Kramer et al., 2004). Additionally, recent work indicates that when parent reports about pre-to-post treatment improvements in the child's functioning disagree with the reports of either children or independent laboratory raters, interviewers' impressions of treatment response systematically agree more with parent reports relative to the other informants' reports (De Los Reyes, Alfano et al., 2011). Thus, primary outcome measures are difficult to interpret as representative estimates of treatment efficacy because they are often based on clinicians' reports, and empirical work finds that clinicians systematically favor certain informants' reports over others.

1.4. The primary outcome method devalues the utility of examining patterns of outcome effects across multiple outcome measures

Lastly, despite the availability of multiple outcome measures within a controlled trial (i.e., primary and secondary outcomes), the primary outcome method downplays the potential utility of examining patterns of multiple outcomes identified within studies and between studies of a particular intervention. Indeed, if the primary outcome measure supports a treatment's efficacy, it logically follows that an investigator is well within his or her right to ignore the extent to which the secondary outcome measures yielded similar conclusions. Conversely, a scenario in which a primary outcome measure fails to support the treatment's efficacy logically results in an investigator (theoretically) being forced to conclude that the evidence does not support the treatment's efficacy; regardless of what the secondary outcome measures indicate.

The realities of evidentiary interpretations under the primary outcome method lie in stark contrast to the reasons that researchers often have for collecting multiple outcome measures. For instance, in clinical child assessments researchers often use multiple informants because informants systematically vary in how or under what circumstances they observe the children being assessed (De Los Reyes, 2011). That is, researchers often collect information from parents and teachers because one informant primarily observes children in the home setting (parent) and the other in a school setting (teacher) (Kraemer et al., 2003). As such, when differences arise between these reports, researchers may have at their disposal a window into how children's behavior varies across situations. For example, if two informants' reports disagreed on whether they indicated that treatment improved a child's functioning, this may reflect improvements in one, both, or none of the settings in which the informants primarily observed the child (e.g., treatment *worked* in reducing problems at school and not home; De Los Reyes & Kazdin, 2008).

A number of recent studies in the clinical child literature support the idea that inconsistent findings based on multiple outcome reports may yield important information on the nature and extent of a treatment's efficacy. For example, two recent studies indicate that measurements of the discrepancies between informants' reports are stable both within a single intake clinical assessment as well as when assessed before and after treatments administered within a controlled trials setting (De Los Reyes, Alfano, & Beidel, 2010; De Los Reyes, Youngstrom et al., 2011). Further, two recent studies indicate that the extent to which parents and teachers provide discrepant reports of children's aggressive and oppositional behavior relates to differences in the situations in which parents and teachers observe children expressing these behaviors (De Los Reyes, Henry, Tolan, & Wakschlag, 2009; Hartley, Zakriski, & Wright, 2011). Consistent with this work, a recent meta-analytic review of controlled trials testing psychological treatments for childhood anxiety and conduct problems suggests that investigators can identify patterns of outcome effects within studies using multiple outcome measures, and in some circumstances, these patterns systematically relate to who provided the outcome reports (De Los Reyes & Kazdin, 2009). In sum, the primary outcome method delegitimizes the potentially invaluable practice of identifying how or under what circumstances different outcome measures within a controlled trial yield supportive evidence of treatment efficacy.

1.5. Summary comments

We have highlighted a number of limitations of the primary outcome method for gauging treatment response within controlled trials. Indeed, these limitations are symptoms of a significant problem in controlled trials research. Specifically, in the absence of definitive *gold standard* outcome measures and the availability of multiple reliable and valid outcome measures, investigators have nonetheless designated specific measures as gold standards without the basis for doing so. This practice directly contradicts recent work suggesting primary outcome measures: (a) that reveal null findings are often ignored in favor of alternative evidence that points to a treatment's efficacy; (b) often reveal evidence that is inconsistent with evidence based on other outcome measures; (c) might not represent multiple informants' reliable and valid perspectives of treatment response; and (d) downplay the information that might be gained from studying patterns of multiple outcome effects observed within controlled trials.

In light of these limitations, it is important to note that not all controlled trials have applied the primary outcome method to test efficacy (for reviews see De Los Reyes & Kazdin, 2006; Weisz et al., 2005). For example, some notable exceptions include the well-conducted multi-site trial, the Multimodal Treatment Study of Children with Attention-Deficit/Hyperactivity Disorder (MTA), in which researchers subjected multiple measures within a comprehensive baseline assessment to a principal components analysis, and identified a multi-domain outcome battery through which to test efficacy (MTA, 1999). The result of such an outcome battery is that researchers can subject the battery to tests of patterns of outcome findings across the domains assessed and measures used in the battery.

It is also important to note that existing approaches that seemingly address issues raised by the primary outcome method suffer from the same or similar limitations as the primary outcome method. For instance, one might surmise that information taken from multiple outcome measures could be combined using systematic algorithms. Two such algorithms are the "and/or rules." Within the context of controlled trials research, an "and" rule requires at least two (or all) measures to suggest the treatment was efficacious for one to identify the treatment as efficacious (Offord et al., 1996). Conversely, an "or" rule requires only one measure to support the treatment's efficacy (Piacentini, Cohen, & Cohen, 1992; Youngstrom, Findling, & Calabrese, 2003). This approach might be used when researchers anticipate that

they will observe low agreement among outcome measures (Goodman et al., 1997; Lofthouse, Fristad, Splaingard, & Kelleher, 2007). However, this approach is limited for two reasons. First, use of and/or rules is not any more incrementally reliable than interpreting information from the multiple measures independent from one another (Offord et al., 1996). Second, when researchers rely on and/or rules they can fail to detect relations between the assessed behavior and other constructs (e.g., moderators of treatment response) that are identified when using the individual measures (Gizer et al., 2008; Offord et al., 1996; Rubio-Stipec, Fitzmaurice, Murphy, & Walker, 2003). This may occur because often outcome measures are completed by multiple informants who systematically vary in the contexts in which they observe the behaviors being targeted by the treatment (e.g., parents observe behaviors expressed at home whereas teachers observe behaviors expressed at school; De Los Reyes, 2011). Therefore, use of these rules results in losses of opportunities to identify situation-specific instances in which the assessed behavior occurs, and thus the specific instances in which the treatment may have been particularly effective (see De Los Reyes & Kazdin, 2006, 2008). Thus, use of and/or rules may often result in a loss of information about behaviors assessed at outcome.

Another approach that researchers might surmise accounts for the limitations of the primary outcome method is structural equation modeling. Structural equation modeling comprises a set of statistical techniques in which multiple measures developed to assess the same construct are examined in combination to arrive at an unobserved or "latent" representation of the assessed construct (Borsboom, 2005). In controlled trials research, this approach can be used to extract and examine the common variance shared by multiple measures used to assess treatment outcome, particularly those measures that assess outcomes on the same domain (e.g., depressive symptoms). Similar to and/or rules, studies using structural equation modeling often treat the variance not shared by multiple measures as measurement error (for a review see Holmbeck, Li, Schurman, Friedman, & Coakley, 2002). In fact, when researchers discover large discrepancies among measures within a structural model, they may interpret these discrepancies as reflecting low measurement reliability (e.g., Arseneault et al., 2003; Zhou, Lengua, & Wang, 2009). However, these interpretations are inconsistent with work reviewed previously indicating that differences among measures or unique variance may reveal important information about assessed behaviors. Therefore, the limitations of the primary outcome method and absence of existing approaches that adequately address its limitations point to the need to develop a new method for assessing efficacy within controlled trials testing treatments for mental disorders.

2. A research agenda for developing a replacement for the primary outcome method

We present an agenda to guide future research and theory on the development of an approach to replace the primary outcome method for testing efficacy within controlled trials. Specifically, we discuss the guiding principles of the agenda, its main components, and the challenges to enacting it in future research.

2.1. Principles guiding the research agenda

2.1.1. To date, no gold standard outcome measure exists to test the efficacy of any treatment, for any diagnostic condition, for any treatment population

The key principle guiding the research agenda described below – and that from which the subsequent principles logically follow – is that no definitive outcome measure exists through which to gauge treatment response for any one intervention. As mentioned previously, this is a well-accepted principle of clinical assessment, broadly construed (Hunsley & Mash, 2007). Thus, it is completely reasonable to assume that this principle holds for testing efficacy within controlled trials.

2.1.2. In the absence of a gold standard, multiple reliable and valid outcome measures should be administered in every controlled trial, with each measure given equal weight in testing efficacy

Consistent with the *no gold standard* principle, it logically follows that if no definitive outcome measure exists, and multiple reliable and valid measures exist to gauge treatment response, then multiple reliable and valid outcome measures should be administered within trials to test efficacy.[1] Relatedly, if no definitive measure exists, then one can reasonably conclude that no definitive method exists by which to determine if any one of multiple outcome measures used in a trial is the *best* measure. Thus, another principle guiding the research agenda is that within a single controlled trial, one should equally weight the evidence gathered from the multiple outcomes administered within the trial. For example, in a trial testing a treatment for adulthood major depressive disorder, findings based on one reliable and valid depressive symptom outcome measure out of the six reliable and valid depressive symptom measures cannot be viewed as any more definitive for gauging efficacy than any of the remaining five. Here, this *equal weighting* criterion assumes that investigators will (a) apply such a criterion to testing efficacy using reliable and valid outcome measures for the treatment being tested (i.e., unreliable and/or invalid measures would not be included in the weighting); (b) apply statistical corrections for conducting multiple tests of treatment efficacy within the trial; and (c) equally weight measures that one could reasonably apply the same hypothesis when evaluating the measures (e.g., investigators equally weighting findings based on four social anxiety symptom measures, within a trial testing treatments for social anxiety disorder).

2.1.3. One clarifies the interpretive power of giving equal weight to multiple outcome measures by examining the extent to which evidence supporting efficacy replicates across findings based on these measures

When invoking the principle of equal weight for multiple pieces of evidence gathered within a controlled trial, it is important to clarify how one will interpret this evidence. Indeed, without such a clarification it is likely that investigators would revert to using the primary outcome method or other methods that suffer from the same or similar limitations (see De Los Reyes & Kazdin, 2006). Fortunately, one can apply well-accepted measurement principles to clarify the interpretation of findings within controlled trials.

Specifically, the likely consensus among controlled trial researchers is that what greatly facilitates the interpretability of controlled trial findings is when two or more controlled trials testing the efficacy of the same treatment, for outcomes on the same diagnostic condition, and in the same or similar treatment populations yield the same or similar conclusions (Chambless & Ollendick, 2001; Kazdin, 2003). That is, given two tests of a hypothesis examined under similar conditions, the first test ought to yield findings that can be *replicated* by the findings derived from the second test, if one wishes to lend credibility to the veracity of the findings derived from any one of the two tests (e.g., Hempel, 1966). Researchers usually apply this principle to interpreting findings from two independently conducted trials. We hold that researchers should also apply this principle to the multiple outcomes administered *within* a controlled trial. Indeed, two separate investigations of the same treatment are often conducted under disparate circumstances (e.g., different

[1] The number of outcome measures one uses within a controlled trial should be based on such factors as the nature of the condition being treated, the goals of the treatment package, and how many reliable and valid outcome measures are available to assess efficacy. To our knowledge, no definitive criterion exists by which to determine exactly how many measures should be administered within any one controlled trial to test the efficacy of any one treatment.

investigators, participant characteristics, outcome measures, sample sizes, and statistical power parameters and data analytic plans). In fact, one might reasonably argue that the replication principle is *more applicable* to the case of multiple outcome measures within one controlled trial as opposed to findings compared between two or more trials. This is because many more of the factors noted previously that might differentiate the nature of tests of efficacy as conducted between two or more trials of one treatment are held constant across outcome measures within a single trial of that treatment. Importantly, it is in this consistency in methodological factors across tests of efficacy within a controlled trial that would make one more confident (relative to tests of treatments between two or more trials) that any differences across the tests reflect important variation in the nature of treatment efficacy (Kazdin, 2003). Therefore, the research agenda illustrated below is informed by the importance of examining whether multiple outcome measures within a controlled trial yield replicable findings.[2]

Related to this core principle guiding the research agenda we outline below is that how often findings from multiple outcomes replicate within a controlled trial might *mean* different things depending on the characteristics of the outcome measures and the treatment literature within which the controlled trial resides. For example, multiple informants' reports of children's behavior typically agree more when the informants observe the child's behavior in the same setting (e.g., both parents at home; and two teachers in the same school) and when the behavior being assessed is overtly versus covertly expressed (e.g., aggressive and oppositional behavior versus obsessive thoughts) (Achenbach, 2006). Additionally, these aspects of outcome measures administered within controlled trials may be a powerful tool for understanding whether interventions exact effects under some circumstances (e.g., home-based on spouse's report of patient's mood) and not others (e.g., work-based on report of work productivity) (see De Los Reyes & Kazdin, 2008).

Therefore, multiple factors may affect the extent to which measures yield consistent findings. It follows that the nature of replication across outcome measure findings should not be evaluated in an absolute sense. It would be foolhardy for all treatment literatures to strive to carry out controlled trials in which 90% of all outcome measures yield evidence supportive of the treatment's efficacy. For some treatment literatures, such a replication rate might be possible given the nature of how patients respond to the treatments, requisite levels of statistical power in the trials, and reliability and validity of the outcome measures. For others, such a rate might be impossible. Additionally, lower rates of replication within controlled trials of a treatment might not be a *bad thing* because these rates may reflect that there are particular circumstances in which the treatment *works*—not necessarily that the treatment is ineffective. Thus, the research agenda below is informed by the idea that efficacy and specifically the extent to which multiple outcomes within a controlled trial yield replicable findings should be evaluated relative to the nature of the evidence and the treatment literature within which the trial resides.

3. Research agenda

Consistent with the basic principles outlined previously, the research agenda discussed below should be viewed as a set of recommendations for future research. The first recommendation is that we must understand the rate at which findings based on outcome measures administered within controlled trials replicate or yield similar conclusions. The second is that once we identify these *replication rates*, we must create standardized estimates of these rates within specific treatment literatures. Third, controlled trials research would benefit from the creation of standards by which investigators identify *a priori* the specific outcome batteries they will use within controlled trials that they plan to conduct. Finally, we advocate for the development of a revised outcome evaluation process in which findings observed in a controlled trial are evaluated relative to (a) findings observed in previous controlled trials testing the same or similar treatments and (b) the characteristics of the outcome measures used in the trial.

3.1. Gauge the replication rate of findings based on multiple outcome measures administered within controlled trials

We recommend that investigators conduct quantitative research reviews to assess the rates by which multiple outcome measures within prior trials have yielded supportive evidence of the efficacy of treatments. Specifically, researchers should evaluate how often outcome measures outperform control or comparison conditions using an agreed-upon metric (e.g., p value<0.05; minimum *small* effect size difference as outlined by Cohen, 1988; see Footnote 2). Further, such replication rates should be calculated within studies evaluating the efficacy of the same treatments (or multiple treatments using highly similar techniques or "active ingredients") for the same diagnostic condition and treatment population. By identifying these rates, investigators may build a foundation for gauging the rates of any one trial relative to the rates observed in other trials of the same or similar treatments.

3.2. Identify and construct outcome replication norms within specific treatment literatures

Multiple scientific disciplines have enjoyed a long history of standardized assessment methods. That is, investigators assess patients by administering widely used measures and comparing patients' scores to scores obtained from representative patient samples that completed the same measures (see Groth-Marnat, 2009). For instance, the Wechsler Intelligence Scales for both adults and children include published standardized scores (i.e., *norms*) for representative samples of the general population as well as representative samples of specific populations identified using independent criteria (e.g., learning and intellectual disabilities; ADHD; and ethnic minorities; see Wechsler, 2008a, 2008b). With these norms, assessors can evaluate an individual patient's scores and determine (a) the patient's level of intelligence relative to a given reference group and (b) in the case of diagnostic testing, whether the patient's pattern of scores *matches* the patterns of scores expected from a member of a particular diagnostic group (e.g., adults diagnosed with a learning disability; children diagnosed with ADHD).

Using methods similar to those used in standardized intelligence assessments, we recommend that investigators create *Standardized Replication Rates* (SRRs) as observed within specific controlled trials. That is, within specific treatment literatures (e.g., cognitive therapy

[2] In discussing the importance of examining replicable effects, a crucial issue involves methods of examining replication. That is, one method by which investigators could assess the extent of replicated treatment outcome effects could be the statistical results of hypothesis tests (e.g., whether the test statistics of two or more findings within a single study fall below an *a priori* threshold of statistical significance). This method might be defensible in circumstances in which it is clear that methodological and statistical characteristics that might influence the detection of significant effects were held constant across outcome measures (e.g., missing data, sample size, outcome measure reliability and validity). Indeed, it is often used to conduct systematic qualitative reviews of treatment literatures. However, other methods could be used as well such as identifying replicable magnitudes of treatment effects or effect sizes (see Cohen, 1988; Rosenthal & DiMatteo, 2001). This is because when controlled trials yield variable patterns of statistically significant outcome effects within and between studies, they also often yield variable patterns in observed magnitudes of outcome effects within and between studies (De Los Reyes & Kazdin, 2006). Further, it is possible to assess rates of replication using either of these methods (for examples see De Los Reyes & Kazdin, 2009). Thus, examining how treatment outcome effects replicate across outcome measures administered within and between controlled trials can be done based on the results of both null hypothesis significance testing and effect size calculations, because investigators often identify variable outcomes using either method.

for adult depression; and parent training for childhood ADHD), investigators could standardize or norm-reference observed replication rates of multiple outcomes administered within controlled trials testing the same or similar treatments. In constructing these SRRs within controlled trials, by definition, one would be taking a measure of the variability of outcome findings within controlled trials. Additionally, one can *norm* these within-trial SRRs in relation to a subset of any of a number of salient treatment and study design characteristics including: (a) treatment type; (b) trial sample size; (c) number of outcome measures; (d) targeted diagnostic condition; (e) patient demographics; and (f) comparison condition (e.g., placebo, waitlist, and alternative treatment). For example, one can create separate norms for SRRs observed for controlled trials in a treatment literature, based on whether the treatment tested in a given trial was a psychosocial versus pharmacological treatment. This application of the norming methodology is quite similar to use of norming methods to identify variations in scores indicative of clinically relevant patient dysfunction, depending on the individual patient's gender, such as that seen in widely used behavior checklists (e.g., Achenbach & Rescorla, 2001).

To be clear, we are not advocating an approach in which investigators create a separate norm for each unique combination of treatment and design characteristics found in controlled trials. We are simply arguing that if investigators wish to evaluate outcome replication rates for a trial relative to rates observed for trials conducted under similar circumstances, there are methods available to make these relative evaluations. Investigators can create and apply norming methods to assess the functioning of individual patients. Thus, little keeps investigators from applying these same approaches for norming the evidence used to assess treatment efficacy.

Additionally, there is precedent for applying norming procedures to controlled trial outcomes not only within standardized assessments of individual patient's behaviors (e.g., intelligence, emotional and behavioral problems). Indeed, one version of standardizing the interpretation of outcome evidence is the application of quantitative review methods to controlled trials. Specifically, when investigators examine the evidence gathered across studies gauging the efficacy of a given treatment, they calculate effect sizes or standardized estimates of the magnitudes of the differences between the examined treatment and control or comparison conditions (see Rosenthal & DiMatteo, 2001). Importantly, investigators of treatments for both adults and children often interpret these average effects relative to Cohen's (1988) effect size conventions of what constitutes a *small*, *medium*, or *large* effect for a given effect size metric (e.g., Cohen's *d*; see Cuijpers, Li, Hofman, & Andersson, 2010; De Los Reyes & Kazdin, 2009; Weisz, Jensen Doss, & Hawley, 2006; Weisz, McCarty, & Valeri, 2006). In this way, efficacy is not evaluated absolutely (i.e., size of differences in the absence of a reference point), but relative to what investigators have accepted as the standard for *noteworthy* differences between conditions.

The SRR Approach we outline here is but an extension of a practice already conducted in quantitative reviews of the controlled trials literature. The key advance lies in how investigators calibrate the *noteworthiness* of controlled trial outcomes. Specifically, as noted previously, researchers often interpret the average effects observed within a treatment literature based on Cohen's (1988) effect size conventions. There are two differences between these average effects and the SRR Approach. The first is the obvious difference that interpretations of average effect sizes likely do not capture the wide variability in outcome evidence observed within controlled trials (De Los Reyes & Kazdin, 2009). The second is that these average effect size interpretations are also made without reference to the specific treatment or treatments evaluated within the trials. That is, according to Cohen (1988) *small*, *medium*, and *large* effects are interpreted essentially the same way in relation to effect size metrics (e.g., $d = 0.2$, 0.5, and 0.8, respectively). This is regardless of whether one is evaluating evidence from controlled trials of, say, parent training for childhood ADHD or one-session treatment for specific phobia in adults (cf. Weisz, Hawley, & Jensen Doss, 2004; Zlomke & Davis, 2008).

One might argue that creating these norms would constitute an insurmountable task, particularly because many controlled trials would have to have been conducted for treatments of the same clinical condition. However, to date thousands of randomized controlled trials of mental health treatments have been conducted; so much so that investigators carry out *reviews of the reviews* so that researchers studying specific treatments can remain current with the findings (Kazdin, 2008). Thus, in constructing SRRs within controlled trials, one can examine the nature of replication across trials within a given literature and use norm-referencing methods to understand what should be expected of evidence gathered in a typical trial in that literature.

3.3. Advanced registration of outcome batteries used within specific controlled trials

Developing and implementing the SRR Approach described previously places the onus on investigators to uphold sound outcome reporting standards. We recommend that researchers take two standards under heavy consideration. The first is to continue collecting multiple reliable and valid outcome measures to evaluate efficacy within controlled trials. The second is that, for any one controlled trial, researchers should only test efficacy via outcome measures that were selected in advance of conducting the trial. Indeed, otherwise performing controlled trials might result in implementation of a multiple outcome method that suffers from the same key limitations as that of the primary outcome method (see Boutron et al., 2010). Thus, to facilitate the successful implementation of sound outcome reporting standards consistent with this research agenda, investigators ought to engage in similar outcome registration practices as those currently required of investigators reporting primary and secondary outcomes.

Specifically, recall that databases exist within which investigators register the outcomes used within controlled trials (see De Angelis et al., 2004). Similarly, we recommend that a database exist through which, in advance of conducting controlled trials, researchers register the multiple outcome battery they will implement to test efficacy within the trial. Ideally, this database should include the characteristics through which these outcomes are expected to be evaluated (e.g., comparison condition, sample characteristics and proposed sample size, and replication rate norms of the larger treatment literature). In short, by requiring the *a priori* registration of the outcome batteries used within controlled trials, investigators will be held accountable to the specific outcomes they proposed to use in advance of the trial.

3.4. Evaluate outcome evidence in relation to replication norms and the characteristics of the evidence that supports (and/or fails to support) the efficacy of the treatment(s) examined

The culmination of the three components of the proposed research agenda is a fundamentally new approach toward understanding the efficacy of treatments as tested within controlled trials. Relative to the primary outcome method, two fundamental differences are readily apparent. First, the research agenda may guide researchers toward conducting quantitative reviews to identify how often the multiple outcomes tested within controlled trials yield replicable evidence supporting the efficacy of the treatments examined. By doing so, researchers may calculate SRRs as observed within controlled trials. In turn, these SRRs would allow researchers to interpret the multiple outcome findings they observed within their controlled trial of a treatment relative to the multiple outcome findings observed within

other trials of the same or similar treatments (e.g., similar treatment, diagnostic condition, treatment population, and sample and design characteristics).

Second, the research agenda places increased focus on interpreting outcome findings based on how the outcome information was collected. That is, not only should outcome evidence within a trial be evaluated relative to evidence gathered in other trials of the same treatment, the characteristics of the outcome evidence should also guide interpretations of a treatment's efficacy. Therefore, relative to using the primary outcome method for gauging efficacy, the proposed agenda would facilitate comparative interpretations of controlled trial outcomes and provide a more parsimonious account of the reasons why differences might arise among outcome measures both within trials and between trials of the same treatment.

4. Challenges posed by research agenda

It is important to highlight challenges to implementing the proposed research agenda.

4.1. Comparison conditions for gauging replication rates

First, when attempting to measure and interpret outcome replication rates within controlled trials, the question arises as to how such rates will be constructed relative to use of the controlled trials design. Indeed, findings are the result of statistical comparisons between the outcomes of participants randomly assigned to receiving either the treatment or treatments under investigation or some other condition (Kazdin, 2003). Needless to say, treatments can be compared to what are hypothesized to be relatively inert conditions such as pill placebos or waitlist control conditions as well as alternative active treatment conditions. Thus, when constructing replication rates of findings within a controlled trial, to what conditions will treatments be compared and findings derived?

The answer to the question likely depends on the research question being addressed within the controlled trials examined. For example, if within a treatment literature the basic question asked by multiple trials is, *Is this treatment efficacious?*, then outcome replication rates might be constructed based on tests of the treatment compared to relatively inert conditions (e.g., extent to which the treatment outperforms a simple expectation of improvement [pill placebo]). Alternatively, if the research question has progressed beyond that of efficacy relative to inert conditions and toward a question such as, *Does the treatment work as well or better than alternative treatments for the same condition?*, then rates should be constructed based on comparisons between the treatment and active treatment conditions. In any event, we encourage researchers to compare outcome replication rates within trials testing the same treatments, and among those trials testing efficacy relative to the same or similar comparison conditions.

4.2. Feasibility of administering multiple outcome measures within one controlled trial

A second challenge to implementing the proposed research agenda may prove more practical. That is, to what extent can controlled trials within specific treatment literatures test the multiple reliable and valid outcome measures to which they have access? Indeed, with the increased availability of outcome findings generated from large multi-site controlled trials, future trials may have to be conducted on similarly large scales. Thus, the current state of controlled trials research may reduce the ability of any one study to implement a comprehensive outcome battery, particularly if assessments must be coordinated across multiple sites.

An issue similar to that of the multi-site treatment design has arisen in studies in other literatures. As an example, consider research testing the presence of gene (e.g., the 5-HTT serotonin transporter gene) × environment interactions and their ability to predict maladaptive stress responses and depressive symptoms in humans. In this work, often large-scale studies have sample sizes that preclude implementing comprehensive assessments of the key environmental risk examined in these studies: exposure to stressful life circumstances (Caspi, Hariri, Holmes, Uher, & Moffitt, 2010). This creates variability across studies in both sample size and the quality of measurement of environmental risk, and as such may create a spurious negative association between sample size and measurement quality (Caspi et al., 2010).

The issue of sample size and measurement quality has significant implications for the study of gene × environment interaction effects. This is because, as Caspi et al. (2010) note, when calculating the average effect size observed across studies, quantitative reviewers often weight a given study's effects by the study's sample size. Stated another way, quantitative reviewers that allow sample size to differentially influence their calculations of average effect sizes might inadvertently allow average effect sizes to be adversely influenced by studies with poor quality measurement of key variables. Interestingly, many large-scale studies fail to replicate gene × environment interaction effects observed in other relatively smaller studies for which investigators study stressful event exposure comprehensively, reliably, and validly (see Caspi et al., 2010; Monroe & Reid, 2008).

Similar to gene × environment interaction research, it might be that controlled trial investigators will argue it is not feasible to carry out any outcomes method other than the primary outcome method for many controlled trials. However, as demonstrated by the outcomes of similar approaches in the gene × environment interaction literature, in sacrificing measurement quality large-scale studies run the risk of encountering inconsistent and unreliable research findings across studies. This may result in failures to replicate effects across studies and further confusion as to the efficacy of treatments.

4.3. Statistical power and the evaluation of multiple outcomes

One final challenge to address is that implementing our proposed agenda would introduce statistical power issues that come from evaluating multiple outcome measures within a study. That is, examining how often multiple findings replicate within a study would introduce the challenge of not only deciphering the statistical power needed to detect a given effect on one measure, but also the statistical power needed to observe replicated effects across multiple measures. Indeed, such a challenge would be further exacerbated by the fact that many of the measures used within controlled trials often yield findings that are inconsistent with each other (Achenbach, 2006). Thus, it may be the case that attaining estimates of the statistical power to observe replicated effects would have to incorporate a number of other parameters beyond the core parameters of interest in power analyses with a single measure (i.e., sample size, anticipated effect size magnitude, and p value). Specifically, power estimates would have to take into account, among other factors, the number of measures and the expected rate of agreement among them (e.g., how often each of the measures agree on, at minimum, small intervention effects).

At the same time, the purpose of conducting power analyses across multiple outcome measures would not be to determine the requisite power through which to identify significant effects on *all* measures. Indeed, as mentioned previously the rates of agreement among outcome measures within a treatment literature might translate into rates of replicated effects within that literature that fall well under 100%. Yet, identifying replication rates well under 100% would not necessarily indicate that the treatments examined are particularly ineffective. This possibility would introduce the importance of examining whether systematic patterns exist within replication

rates (e.g., variations in the magnitudes of effects across outcome measures in trials, depending on the type of treatment examined in the trials). Additionally, carrying out statistical power analyses for identifying outcome effects across measures would be an important endeavor in and of itself within any treatment study administering more than one outcome measure. This is because such analyses, particularly when conducted *a priori*, would provide researchers with a guide for understanding whether the study design parameters would provide for sufficient power to identify consistent effects across measures. Thus, future research should be dedicated to deciphering statistical power rates for observing replicated effects within trials.

5. Concluding comments

One of the most important questions addressed in mental health research is whether a mental disorder treatment evaluated within a controlled trial outperforms control or comparison conditions on reliable and valid indices of outcome effects. Investigative teams that hold that they have created a substance or technique that successfully ameliorates the impairment caused by a mental disorder are making a claim that they have changed aspects of the human condition that are difficult to assess, let alone treat. By definition, these claims rise to the level of extraordinary claims that require extraordinary evidence (Sagan, 1980). Therefore, reporting that a treatment successfully improves outcomes based on one *nondefinitive* measure when other measures exist and could defensibly be used in its place is *demonstrably unextraordinary*.

Current evidentiary standards as practiced within controlled trials do not allow investigators to collect and draw interpretations from *extraordinary evidence*. In order to improve upon these standards, we must develop new approaches that focus on what scientists value: (a) use of multiple methods and measurement approaches when no single definitive measure exists and (b) assessing how often findings based on one measurement approach replicate findings based on other approaches. As scientists, we can improve our understanding of whether, how, and why our treatments work by renewing and improving our focus on how different outcome indices tell us the same (or different) stories about our treatments.

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.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington: University of Vermont, Research Center for Children, Youth, & Families.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychological Association Interdivisional Task Force on Child and Adolescent Mental Health (2007). Links for finding evidence-based interventions. Retrieved April 10, 2011, from. http://ucoll.fdu.edu/apa/lnksinter.html.

Arseneault, L., Moffitt, T., Caspi, A., Taylor, A., Rijsdijk, F., Jaffee, S., et al. (2003). Strong genetic effects on cross-situational antisocial behaviour among 5-year-old children according to mothers, teachers, examiner-observers, and twins' self-reports. *Journal of Child Psychology and Psychiatry*, *44*, 832–848.

Birmaher, B., Axelson, D. A., Monk, K., Kalas, C., Clark, D. B., Ehmann, M., et al. (2003). Fluoxetine for the treatment of childhood anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 415–423.

Blue Cross and Blue Shield of Texas (BCBSTX) (2007). BCBSTX evidence based measures program. Retrieved April 10, 2011, from. http://www.bcbstx.com/bluecompare/ebm.htm.

Borsboom, D. (2005). *Measuring the mind*. New York: Cambridge University Press.

Boutron, I., Dutton, S., Ravaud, P., & Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *Journal of the American Medical Association*, *303*, 2058–2064.

Bowden, C. L., Calabrese, J. R., McElroy, S. L., Gyulai, L., Wassef, A., Petty, F., et al. (2000). A randomized, placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatients with bipolar I disorder. *Archives of General Psychiatry*, *57*, 481–489.

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

Caspi, A., Hariri, A. R., Holmes, A., Uher, R., & Moffitt, T. E. (2010). Genetic sensitivity to the environment: The case of the serotonin transporter gene and its implications for studying complex diseases and traits. *American Journal of Psychiatry*, *167*, 509–527.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Clark, D. B., Birmaher, B., Axelson, D., Monk, K., Kalas, C., Ehmann, M., et al. (2005). Fluoxetine for the treatment of childhood anxiety disorders: Open-label, long-term extension to a controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 1263–1270.

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

Cuijpers, P., Li, J., Hofman, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review*, *30*, 768–778.

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *351*, 1250–1251.

De Los Reyes, A. (2011). Introduction to the special section. More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *40*, 1–9.

De Los Reyes, A., Alfano, C. A., & Beidel, D. C. (2010). The relations among measurements of informant discrepancies within a multisite trial of treatments for childhood social phobia. *Journal of Abnormal Child Psychology*, *38*, 395–404.

De Los Reyes, A., Alfano, C. A., & Beidel, D. C. (2011). Are clinicians' assessments of improvements in children's functioning "global"? *Journal of Clinical Child and Adolescent Psychology*, *40*, 281–294.

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology*, *37*, 637–652.

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.

De Los Reyes, A., & Kazdin, A. E. (2008). When the evidence says, "Yes, no, and maybe so": Attending to and interpreting inconsistent findings among evidence-based interventions. *Current Directions in Psychological Science*, *17*, 47–51.

De Los Reyes, A., & Kazdin, A. E. (2009). Identifying evidence-based interventions for children and adolescents using the range of possible changes model: A meta-analytic illustration. *Behavior Modification*, *33*, 583–617.

De Los Reyes, A., Youngstrom, E. A., Pabón, S. C., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2011). Internal consistency and associated characteristics of informant discrepancies in clinic referred youths age 11 to 17 years. *Journal of Clinical Child and Adolescent Psychology*, *40*, 36–53.

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., et al. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. *Archives of General Psychiatry*, *48*, 851–855.

Gizer, I., Waldman, I., Abramowitz, A., Barr, C., Feng, Y., Wigg, K., et al. (2008). Relations between multi-informant assessments of ADHD symptoms, DAT1, and DRD4. *Journal of Abnormal Psychology*, *117*, 869–880.

Goodman, S., Lahey, B., Fielding, B., Dulcan, M., Narrow, W., & Regier, D. (1997). Representativeness of clinical samples of youths with mental disorders: A preliminary population-based study. *Journal of Abnormal Psychology*, *106*, 3–14.

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken: Wiley & Sons.

Guy, W. (1976). *ECDEU assessment manual for psychopharmacology*. Washington, DC: DHEW.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*, 56–62.

Hartley, A. G., Zakriski, A. L., & Wright, J. C. (2011). Probing the depths of informant discrepancies: Contextual influences on divergence and convergence. *Journal of Clinical Child and Adolescent Psychology*, *40*, 54–66.

Hawley, K. M., & Weisz, J. R. (2003). Child, parent, and therapist (dis)agreement on target problems in outpatient therapy: The therapist's dilemma and its implications. *Journal of Consulting and Clinical Psychology*, *71*, 62–70.

Hazell, P. L., & Stuart, J. E. (2003). A randomized controlled trial of clonidine added to psychostimulant medication for hyperactive and aggressive children. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 886–894.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ: Prentice-Hall.

Holmbeck, G., Li, S., Schurman, J., Friedman, D., & Coakley, R. (2002). Collecting and managing multisource and multimethod data in studies of pediatric populations. *Journal of Pediatric Psychology*, *27*, 5–18.

Hunsley, J., & Mash, E. J. (2007). Evidence-based Assessment. *Annual Review of Clinical Psychology*, *3*, 29–51.

Joiner, T. E., Walker, R. L., Pettit, J. W., Perez, M., & Cukrowicz, K. C. (2005). Evidence-based assessment of depression in adults. *Psychological Assessment*, *17*, 267–277.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist*, *63*, 146–159.

Knopman, D. S., Knapp, M. J., Gracon, S. I., & Davis, C. S. (1994). The Clinician Interview-Based Impression (CIBI): A clinician's global change rating scale in Alzheimer's disease. *Neurology*, *44*, 2315–2321.

Koenig, K., De Los Reyes, A., Cicchetti, D., Scahill, L., & Klin, A. (2009). Group intervention to promote social skills in school-age children with pervasive developmental disorders: Reconsidering efficacy. *Journal of Autism and Developmental Disorders*, *39*, 1163–1172.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *The American Journal of Psychiatry*, *160*, 1566–1577.

Kramer, T. L., Phillips, S. D., Hargis, M. B., Miller, T. L., Burns, B. J., & Robbins, J. M. (2004). Disagreement between parent and adolescent reports of functional impairment. *Journal of Child Psychology and Psychiatry*, *45*, 248–259.

Lambert, M. J., Hatch, D. R., Kingston, M. D., & Edwards, B. C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology*, *54*, 54–59.

Loeber, R., Green, S. M., & Lahey, B. B. (1990). Mental health professionals' perception of the utility of children, mothers, and teachers as informants of childhood psychopathology. *Journal of Clinical Child Psychology*, *19*, 136–143.

Lofthouse, N., Fristad, M., Splaingard, M., & Kelleher, K. (2007). Parent and child reports of sleep problems associated with early-onset bipolar spectrum disorders. *Journal of Family Psychology*, *21*, 114–123.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Mash, E. J., & Hunsley, J. (Eds.). (2005). *Developing guidelines for the evidence-based assessment of child and adolescent disorders [Special issue]. Journal of Clinical Child and Adolescent Psychology*, *34*.

Monroe, S. M., & Reid, M. W. (2008). Gene–environment interactions in depression research. *Psychological Science*, *19*, 947–956.

Multimodal Treatment Study of Children with Attention-Deficit/Hyperactivity Disorder Cooperative Group (1999). A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder. *Archives of General Psychiatry*, *56*, 1073–1086.

Niederhofer, H., Staffen, W., & Mair, A. (2003). A placebo-controlled study of lofexidine in the treatment of children with tic disorders and attention deficit hyperactivity disorder. *Journal of Psychopharmacology*, *17*, 113–119.

Offord, D., Boyle, M., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., et al. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 1078–1085.

Ogles, B. M., Lambert, M. J., Weight, D. G., & Payne, I. R. (1990). Agoraphobia outcome measurement: A review and meta-analysis. *Psychological Assessment*, *2*, 317–325.

Papakostas, G. I., Mischoulon, D., Shyu, I., Alpert, J. E., & Fava, M. (2010). S-Adenosyl methionine (SAMe) augmentation of serotonin reuptake inhibitors for antidepressant nonresponders with major depressive disorder: A double-blind, randomized clinical trial. *American Journal of Psychiatry*, *167*, 942–948.

Pettinati, H. M., Oslin, D. W., Kampman, K. M., Dundon, W. D., Xie, H., Gallis, T. L., et al. (2010). A double-blind, placebo-controlled trial combining sertraline and naltrexone for treating co-occurring depression and alcohol dependence. *Archives of General Psychiatry*, *167*, 668–675.

Piacentini, J., Cohen, P., & Cohen, J. (1992). Combining discrepant diagnostic information from multiple sources: Are complex algorithms better than simple ones? *Journal of Abnormal Child Psychology*, *20*, 51–63.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82.

Rubio-Stipec, M., Fitzmaurice, G., Murphy, J., & Walker, A. (2003). The use of multiple informants in identifying the risk factors of depressive and disruptive disorders: Are they interchangeable? *Social Psychiatry and Psychiatric Epidemiology*, *38*, 51–58.

Sagan, C. (Writer/Host). (1980, December 14). *Cosmos (12): Encyclopaedia galactica* [Television Broadcast]. Arlington, VA: PBS. Viewed October 14, 2010.

Scahill, L., Riddle, M. A., McSwiggin-Hardin, M., Ort, S. I., King, R. A., Goodman, W. K., et al. (1997). Children's Yale–Brown Obsessive Compulsive Scale: Reliability and validity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*, 844–852.

Shear, M. K., Rucci, P., Williams, J., Frank, E., Grochocinski, V., Bilt, J. V., et al. (2001). Reliability and validity of the Panic Disorder Severity Scale: Replication and extension. *Journal of Psychiatric Research*, *35*, 293–296.

Spearing, M. K., Post, R. M., Leverich, G. S., Brandt, D., & Nolen, W. (1997). Modification of the Clinical Global Impressions (CGI) Scale for use in bipolar illness (BP): The CGI-BP. *Psychiatry Research*, *73*, 159–171.

Thurstone, C., Riggs, P. D., Salomonsen-Sautel, S., & Mikulich-Gilbertson, S. K. (2010). Randomized, controlled trial of atomoxetine for attention-deficit/hyperactivity disorder in adolescents with substance use disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *49*, 573–582.

Wagner, K. D., Berard, R., Stein, M. B., Wetherhold, E., Carpenter, D. J., Perera, P., et al. (2004). A multicenter, randomized, double-blind, placebo-controlled trial of paroxetine in children and adolescents with social anxiety disorder. *Archives of General Psychiatry*, *61*, 1153–1162.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale* (4th ed.). San Antonio: Pearson.

Wechsler, D. (2008b). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio: Pearson.

Weisz, J. R., Hawley, K. M., & Jensen Doss, A. (2004). Empirically tested psychotherapies for youth internalizing and externalizing problems and disorders. *Child and Adolescent Psychiatric Clinics of North America*, *13*, 729–815.

Weisz, J. R., Jensen Doss, A., & Hawley, K. M. (2005). Youth psychotherapy outcome research: A review and critique of the evidence base. *Annual Review of Psychology*, *56*, 337–363.

Weisz, J. R., Jensen Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinic care: A meta-analysis of direct comparisons. *American Psychologist*, *61*, 671–689.

Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. *Psychological Bulletin*, *132*, 132–149.

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity, and sensitivity. *The British Journal of Psychiatry*, *133*, 429–435.

Youngstrom, E., Findling, R., & Calabrese, J. (2003). Who are the comorbid adolescents? Agreement between psychiatric diagnosis, youth, parent, and teacher report. *Journal of Abnormal Child Psychology*, *31*, 231–245.

Zaider, T. I., Heimberg, R. G., Fresco, D. M., Schneier, F. R., & Liebowitz, M. R. (2003). Evaluation of the Clinical Global Impression Scale among individuals with social anxiety disorder. *Psychological Medicine*, *33*, 611–622.

Zhou, Q., Lengua, L., & Wang, Y. (2009). The relations of temperament reactivity and effortful control to children's adjustment problems in China and the United States. *Developmental Psychology*, *45*, 724–739.

Zimmerman, M., McGlinchey, J. B., Posternak, M. A., Friedman, M., Attiullah, N., & Boerescu, D. (2006). How should remission from depression be defined? The depressed patient's perspective. *American Journal of Psychiatry*, *163*, 148–150.

Zlomke, K., & Davis, T. E. (2008). One-session treatment of specific phobias: A detailed description and review of treatment efficacy. *Behavior Therapy*, *39*, 207–223.