2011

Making Sense of Meta-Analysis: A Critique of "Effectiveness of Long-Term Psychodynamic Psychotherapy"

Julia H. Littell
Bryn Mawr College, jlittell@brynmawr.edu

Aron Shlonsky

Recommended Citation
Littell, Julia H., and Aron Shlonsky. "Making Sense of Meta-Analysis: A Critique of 'Effectiveness of Long-Term Psychodynamic Psychotherapy.'" *Clinical Social Work Journal* 39, no. 4 (2011): 340-346, doi: 10.1007/s10615-010-0308-z.

Making Sense of Meta-analysis:

A Critique of "Effectiveness of Long-Term Psychodynamic Psychotherapy"

Julia H. Littell

Graduate School of Social Work and Social Research, Bryn Mawr College

and

Aron Shlonsky

Factor-Inwentash Faculty of Social Work, University of Toronto

2011

Clinical Social Work Journal, 39: 340-346

Running head: Making sense of meta-analysis

Keywords: Evidence-based practice, meta-analysis, systematic review, research review, long-term psychodynamic psychotherapy, AMSTAR

Address correspondence to:

Julia H. Littell, Ph.D., Professor

Graduate School of Social Work and Social Research

Bryn Mawr College, 300 Airdale Rd.

Bryn Mawr, PA, 19010 USA

Email: jlittell@brynmawr.edu

Abstract

Evidence-based practice depends in part on knowledge derived from relevant research. For any given topic, there are likely to be many, potentially relevant studies; a careful appraisal and synthesis of the results of these studies is needed to understand the state of the empirical evidence. Meta-analysis is widely used to combine results of quantitative studies; yet this method is unfamiliar to many people and, as a result, meta-analyses are often uncritically accepted. In this article, we argue that meta-analysis is only one component of a good research synthesis. We critique a recent metaanalysis on the effectiveness of long-term psychodynamic psychotherapy, showing that this metaanalysis failed to meet current standards for the conduct and reporting of systematic research reviews and meta-analyses. We demonstrate the use of AMSTAR, a straightforward tool for assessing the quality of systematic reviews and meta-analyses.

Evidence-based practice depends in no small measure on the identification, critical appraisal, and accurate synthesis of relevant research results. Methods of research synthesis have come a long way since Glass (1976) coined the term "meta-analysis" to describe the statistical combination of data from multiple studies. Meta-analysis is now widely used in the social, behavioral, and health sciences; yet, many people are unfamiliar with this approach and, as a result, meta-analyses are often uncritically accepted. In this article, we critique a recent meta-analysis on the effectiveness of long-term psychodynamic psychotherapy (LTPP) by Leichsenring and Rabung (2008). This meta-analysis appeared in *JAMA*, where Glass himself pronounced that it "was carefully performed" (2008, p. 1589) and others disagreed (Beck & Bhar, 2009; Kriston, Holzel, & Harter, 2009; Glass, 2008; Roepke, 2009; Thombs, Bassel, & Jewett, 2009). We argue that meta-analysis is only one component of a systematic approach to reviewing research, and the LTPP meta-analysis failed to meet current standards for the conduct and reporting of systematic reviews and meta-analyses.

The LTPP meta-analysis is controversial in terms of its content as well as its methods. For many years now, the treatment and research worlds have witnessed the ascendancy of short-term cognitive and behavioral therapies. At face value, the article by Leichsenring and Rabung calls into question the move toward short-term treatment of a range of mental health issues.

As evidence-based practice has developed over the years, so too have the tools needed to quickly and competently evaluate the presentation of findings from different types of studies, including random controlled trials, studies of diagnostic accuracy, observational studies, systematic reviews, and meta-analyses (see http://www.equator-

network.org). There are now also user-friendly tools to evaluate the quality of completed studies. We demonstrate the application of just such a tool (AMSTAR; Shea et al., 2007) for assessing the quality of systematic research reviews.

Background

Systematic reviews and meta-analysis can provide comprehensive, accurate, and useful summaries of empirical research on a wide range of topics that are important to clinicians and consumers. Careful research syntheses can capture what is known about the incidence and prevalence of various conditions; the associations between behavioral, psychological, social, and health conditions; the accuracy of various diagnostic tests; and the effectiveness of treatments.

However, as a "stand alone" method, meta-analysis is entirely insufficient for research reviews, just as statistical analysis is insufficient for survey research. Survey researchers are concerned with sampling, measurement, and data collection, in addition to proper analysis of data. Similarly, the validity of a meta-analysis hinges on a series of decisions that are made during the review process: decisions about the types of studies that will be included, strategies to locate relevant studies, critical appraisal of those studies, and steps to minimize bias and error at each step in the review process.

There is a large body of empirical research that places meta-analysis in its proper context: as a potentially useful set of statistical techniques for research synthesis in the context of a systematic review (Littell, Corcoran, & Pillai, 2008). Systematic reviews, such those produced by the Campbell Collaboration and the Cochrane Collaboration, use transparent and replicable methods to minimize bias and error. Unfortunately, these strategies are absent in many published meta-analyses in psychology, social work,

medicine, and other fields.

The meta-analysis on effectiveness of LTPP by Leichsenring & Rabung (2008) is a case in point: it is one of many recent, published meta-analyses that ignored much of the available evidence about the conduct and reporting of valid research syntheses.

Meta-analysis of the effectiveness of LTPP

Leichsenring and Rabung (L&R; 2008) reported results of a meta-analysis of 11 randomized controlled trials and 12 observational studies of individual psychodynamic psychotherapy. Study participants had chronic, complex, and multiple mental disorders (e.g., anxiety, depression, and personality disorders). In some studies, LTPP was compared with shorter forms of therapy (e.g., inpatient treatment, partial hospitalization, cognitive therapy, dialectic behavioral therapy, or nutritional counseling); in other studies LTPP was evaluated by comparing post-treatment results to baseline data. Measures of effectiveness focused on change over time within the LTPP groups. Meta-analysis was used to combine these results across studies.

L&R claimed that their meta-analysis was conducted in a manner "consistent with recent guidelines for the reporting of meta-analyses" (2008, p. 1552). This is only partially true. They cited two sets of reporting guidelines: the QUOROM statement (Moher et al., 1999; which was subsequently replaced by the PRISMA statement; Moher et al., 2009) and the MOOSE statement (Stroop et al., 2000). We compared their article with the reporting standards they cited and found many gaps in the substance and clarity of their report. We used the Assessment of Multiple SysTemAtic Reviews (AMSTAR) tool (Shea et al., 2007) to assess the methodological quality of the L&R meta-analysis. AMSTAR ratings are shown in Table 1 and discussed below.

1. An a priori design should be provided

The Cochrane and Campbell Collaborations require a detailed protocol for a systematic review that spells out the central questions, objectives, criteria, search strategies, and methods for the review. This protocol is peer-reviewed and is publicly available prior to the commencement of the review. Detailed protocols increase transparency and reduce discretion in the conduct of systematic reviews and meta-analyses. L&R did not provide an *a priori* design for their meta-analysis (i.e., no protocol was made public in advance of the completed review).

The research question should be established before the conduct of the review. The central question in the L&R meta-analysis remains unclear. To understand the effects of LTPP we must make explicit comparisons to other treatments or conditions. Are we interested in the absolute effects of LTPP (compared with no treatment) or relative effects (compared with other treatments)? L&R included some studies that compared LTPP to shorter therapies, alternative treatments, or treatment as usual; but they also included studies that had no comparison or control groups whatsoever. For the latter group of studies, only within-group comparisons (pre and post LTPP) are available. These studies address a fundamentally different question: that is, whether clients who received LTPP demonstrated changes over time. Observed changes could be due to treatment or to many other factors. Absent comparisons to parallel cohorts (differently treated groups assessed in the same timeframe), within-group changes are poor indicators of treatment "effectiveness." Given that the studies included in the L&R meta-analysis addressed different questions, the central question of the meta-analysis is unclear.

Inclusion criteria should be established before the conduct of the review. Criteria for inclusion in the L&R meta-analysis were not entirely clear, and some criteria seem to evolve during the review with post-hoc exclusions of some studies (e.g., those that compared LTPP to inpatient treatment, and studies that included participants who did not have Axis I or II diagnoses). There was no detailed explanation of reasons for study exclusion (also a QUOROM requirement).

2. Study selection and data extraction should be conducted by at least two independent reviewers.

It is unclear how study selection decisions were made in the LTPP meta-analysis. L&R indicated that they used duplicate data extraction (presumably after studies were selected) and a consensus procedure to resolve disagreements for at least some items. However, they also indicated that they used mean study quality ratings instead of a consensus procedure.

3. A comprehensive literature search should be performed.

Three databases were searched, but the authors did not present a clear search strategy. It is not clear how manual searches and contacts with experts were conducted. The search strategy is neither transparent nor replicable (e.g., it is unclear whether electronic searches were limited to keywords associated with articles or covered words that appeared anywhere in the text).

4. *Publication status should not be used as an inclusion criterion.*

Meta-analyses that are limited to published studies tend to over-estimate treatment effects (Hopewell et al., 2007), and it is not clear whether L&R did enough to prevent this.

L&R mentioned publication dates in their inclusion criteria, but also indicated that they considered one unpublished study. Their meta-analysis was limited to published studies.

5. *A list of included and excluded studies should be provided.*

L&R did not list excluded studies, nor did they discuss specific reasons for study exclusion (a stipulation of the QUOROM statement).

6. *Characteristics of included studies should be described.*

L&R provided partial information in the studies included in their review. They did not provide information on participants' demographic characteristics or attrition in the primary studies. L&R did not follow QUOROM guidelines which instruct reviewers to "present simple summary results (for each treatment group in each trial, for each primary outcome); present data needed to calculate effect sizes and confidence intervals in intention-to-treat analyses (e.g., 2x2 tables of counts, means and SDs, proportions)" (Moher et al., 1999, p. 1897).

7. *Scientific quality of included studies should be assessed and documented*

L&R used a simple, linear scale to measure overall study quality. However, overall quality scores conflate distinct methodological issues, which may have disparate effects on the reliability and validity of results (Higgins & Green, 2008; Wells & Littell, 2009). Hence, the "use of summary scores from quality scales is problematic [and] it is preferable to examine the influence of key components of methodological quality individually" (Jüni et al., 2001, p. 42).

L&R state that they only included studies with "reliable" outcome measures, but they do not indicate what types of reliability were considered (e.g., inter-rater, test-retest,

internal consistency?) and what standards were used to determine whether measures were reliable.

8. Scientific quality of included studies should be used appropriately in formulating conclusions.

L&R displayed considerable confusion about research designs and the types of inferences that can be drawn from them. First, they contrast randomized controlled trials (RCTs) with "effectiveness studies," yet many RCTs (field trials) are effectiveness studies. Second, they use the term "overall effectiveness" to refer to changes within groups over time (differences between pre-tests and post-tests); however, as indicated above, within-group differences are measures of change over time, not measures of treatment effectiveness. Studies of treatment effects must control threats to internal validity (e.g., selection bias, maturation, testing effects, other events; see Shadish, Cook & Campbell, 2002). Within-group analyses do not account for the fact that any observed changes could be due to many factors outside of treatment.

For reasons that are entirely unclear, in some analyses L&R eliminated the all-important between-group comparisons in RCTs and focused only on changes within the LTPP group in the RCTs. In essence, they converted the RCTs into observational studies, stripping the RCTs of their ability to support causal inferences about the effects of LTPP! L&R claim that there were no significant differences in the results of RCTs versus observational studies, but this statement is meaningless because the RCTs were essentially converted to observational studies (stripped of their control groups) in this analysis.

L&R stated that they used data on all study participants (intent-to-treat analyses) if these data were available. However, it appears that they combined intent-to-treat samples

in some studies with data from program completers in other studies. The latter are vulnerable to attrition biases, because program dropouts are generally less satisfied with the intervention and have less favorable outcomes than program completers. Further, the post-treatment and follow-up assessments for LTPP groups occurred considerably later than similar assessments for comparison groups in some studies; this may have exacerbated problems with differential attrition. That is, it is much easier for someone to complete a 12-week span of therapy (e.g., CBT) than a much longer course of treatment. Thus, if offered the chance, many completers of short-term therapy might not have finished long-term psychotherapy. Since people who complete long-term therapy probably differ from people who complete short-term therapy (in terms of motivation and other unmeasured factors), differential attrition could explain observed differences between LTPP completers and other groups. L&R provide no discussion of attrition bias, outcome reporting bias, or missing data, although this is required in the QUOROM statement which they cited.

9. Appropriate methods should be used to combine the findings of studies.

L&R use unconventional effect size metrics that are not fully explained (hence, the analysis is neither transparent nor replicable). Bhar and colleagues (2010) show that the approach taken by L&R produces "grossly inaccurate" and inflated estimates of effects. L&R followed 20 year-old recommendations on the statistical synthesis of data from multiple studies. For RCTs, they should have used standardized mean differences with adjustments for small sample bias (Hedges' *g*) (see Lipsey & Wilson, 2001; Higgins & Green, 2008).

L&R did conduct homogeneity tests and used random-effects models, as appropriate.

In some analyses, L&R made the mistake of combining measures of conceptually distinct outcomes. Because treatments can have positive effects on some outcomes and null or negative effects on others, the combination of conceptually distinct outcomes produces meaningless overall estimates of effects.

To avoid "fishing" for significant effects, subgroup analyses should be limited to a few *a priori* contrasts (Higgins & Green, 2008). L&R conducted many subgroup analyses, unspecified beforehand, thus inflating the risk of a type I error.

10. *The likelihood of publication bias should be assessed.*

L&R used outdated and inadequate methods to assess the likelihood of publication bias. The file drawer number (or failsafe N) is misleading and should never be used (Becker, 2005). L&R also used a rank correlation test published in 1994, although many newer and better methods were available (Rothman, Sutton, & Bornstein, 2005).

11. Conflicts of interest and sources of support should be clearly acknowledged.

As noted in JAMA, L&R provided no financial disclosures. Leichsenring is the first author on one of the studies in this review and several previously published meta-analysis of psychodynamic psychotherapy.

Summary and conclusions

Credible research syntheses use transparent, replicable methods and attempt to minimize well-known sources and types of bias and error at each step in the review process. Guidelines for the conduct and reporting of systematic reviews and meta-analyses are well developed (see Higgins & Green, 2008; Moher et al., 2009). These guidelines are based on a large body of methodological research on the conduct of primary studies and research syntheses. Unfortunately, many published meta-analyses and other types of

research reviews do not take this body of knowledge into account and do not follow established guidelines for the conduct and reporting of research reviews.

The AMSTAR tool provides a simple way to assess whether central issues were adequately addressed in a research review. The meta-analysis by L&R fell short on all 11 items on the AMSTAR instrument. This suggests that the L&R meta-analysis of LTPP is neither transparent nor replicable, and it lacks adequate controls for bias and error. As such, the meta-analysis by L&R is not a credible synthesis of research on the effectiveness of LTPP; thus, it does not provide convincing evidence for or against the use of LTPP or other treatments.

Chalmers and colleagues (1993) noted that the failure to use scientific methods in locating, appraising, and synthesizing research can have serious consequences, because service providers rely more on research reviews than primary studies to learn about treatment effects. “Thus, failure by reviewers to apply scientific principles to this secondary research can have adverse consequences for patients” and can lead to wasted service resources (Chalmers et al., 1993, p. 412).

The information age will continue to push the creation and availability of new knowledge, and credible syntheses of this knowledge will become increasingly important. As a method of research synthesis, meta-analysis is appealing in many respects. Meta-analysis offers a powerful set of statistical tools for combining and analyzing results across studies. Compared with narrative reviews of research, meta-analysis provides more accurate summaries of quantitative studies (Bushman & Wells, 2001). However, like any statistical technique, meta-analysis can be misused and its results can be misleading. Just as in survey research, statistical analysis may be necessary but insufficient for valid results.

The validity of a meta-analysis depends largely on whether it is conducted in the context of a comprehensive and rigorous systematic review. AMSTAR and other independent and easy-to-use tools can be useful for practitioners and policy-makers in their efforts to evaluate the quality of seemingly complex studies and research reviews.

References

Beck, A. T., & Bhar, S. (2009). Analyzing the effectiveness of long-term psychodynamic psychotherapy. *JAMA, 301,* 931.

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Bornstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustment* (pp. 111-125). West Sussex, England: John Wiley & Sons, Ltd.

Bhar, S. S., Thombs, B. D., Pignotti, M., Bassel, M., Jewett, L., Coyne, J. C., & Beck, A. T. (2010). Is longer-term psychodynamic psychotherapy more effective than shorter-term therapies? Review and critique of the evidence. *Psychotherapy and Psychosomatics*, *79*, 208-216.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personal and Social Psychology Bulletin*, *27*, 1123-1130.

Chalmers, I., Enkin, M., & Keirse, M. J. N. C. (1993). Preparing and updating systematic reviews of randomized controlled trials of health care. *Milbank Quarterly, 71,* 411-437.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Glass, R. M. (2008). Psychodynamic psychotherapy and research evidence: Bambi survives Godzilla? *JAMA, 300,* 1587-1589.

Higgins, J. P. T., & Green, S. (Eds.). (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, Ltd. Available at: http://www.cochrane.dk/cochrane/handbook/hbook.htm

Hopewell, S., McDonald, S., Clarke, M., & Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *The Cochrane Database of Systematic Reviews*, *2007*(Issue 2), Art. No.: MR000010. doi:10.1002/14651858.MR000010.pub3

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *British Medical Journal*, *323*(7303), 42-46.

Kriston, L., Holzel, L., & Harter, M. (2009). Analyzing the effectiveness of long-term psychodynamic psychotherapy. *JAMA, 301,* 930-931.

Leichsenring, F., & Rabung, S. (2009). Analyzing the effectiveness of long-term psychodynamic psychotherapy. *JAMA, 301,* 932-933.

Leichsenring, F., & Rabung, S. (2008). The effectiveness of psychodynamic psychotherapy: A meta-analysis. *JAMA, 300,* 1551-1565.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York: Oxford University Press.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. F., & The QUOROM Group. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *The Lancet*, *354*, 1896-1900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*, e1000097. doi:10.1371/journal.pmed.1000097

Roepke, S. (2009). Analyzing the effectiveness of long-term psychodynamic psychotherapy. *JAMA, 301,* 931-932.

Rothstein, H., Sutton, A. J., & Bornstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: John Wiley & Sons Ltd.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., et al. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*. Available at: http://www.biomedcentral.com/1471-2288/7/10

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*, *283*, 2008-.

Thombs, B. D., Bassel, M., & Jewett, L. R. (2009). Analyzing the effectiveness of long-term psychodynamic psychotherapy. *JAMA, 301,* 930.

Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice*, *19*, 52-62.

Table 1: AMSTAR ratings for Leischsenring and Rabung (2008)

AMSTAR item (Shea et al., 2007)	Rating	Notes
1. Was an 'a priori' design provided? The research question and inclusion criteria should be established before the conduct of the review.	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	There is no public protocol for this review; it appears that inclusion criteria were modified during the review.
2. Was there duplicate study selection and data extraction? There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.	☐ Yes ☐ No ☑ Can't answer ☐ Not applicable	Duplicate study selection is not mentioned. Duplicate data extraction was performed.
3. Was a comprehensive literature search performed? At least two electronic sources should be searched. The report must include years and databases used (e.g. Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.	☐ Yes ☐ No ☑ Can't answer ☐ Not applicable	Electronic searches are not described in sufficient detail. Little information is available on search for relevant unpublished studies.

4. Was the status of publication (i.e. grey literature) used as an inclusion criterion? The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.	☐ Yes ☐ No ☑ Can't answer ☐ Not applicable	Only published studies were included.
5. Was a list of studies (included and excluded) provided? A list of included and excluded studies should be provided.	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	List of excluded studies is missing.
6. Were the characteristics of the included studies provided? In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the studies analyzed e.g. age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.	☐ Yes ☐ No ☑ Can't answer ☐ Not applicable	Missing data on demographics and some outcomes.

7. Was the scientific quality of the included studies assessed and documented? 'A priori' methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	Authors adapted and used an overall scale that does not assess risk of bias (selection, performance, detection, attrition bias, etc.). Quality ratings were not provided.
8. Was the scientific quality of the included studies used appropriately in formulating conclusions? The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	Insufficient assessment of methodological quality.
9. Were the methods used to combine the findings of studies appropriate? For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e. Chi-squared test for homogeneity, I^2). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e. is it sensible to combine?).	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	Inappropriate statistical methods.

10. Was the likelihood of publication bias assessed? An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test).	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	Outdated and inappropriate methods for assessment of publication bias.
11. Was the conflict of interest stated? Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.	☐ Yes ☑ No ☐ Can't answer ☐ Not applicable	The authors did not provide a statement on conflict of interest. JAMA noted that authors made no financial disclosures.