Special Article

Psychotherapy and Psychosomatics

Psychother Psychosom 2010;79:208–216
DOI: 10.1159/000313689

Received: May 18, 2009
Accepted after revision: January 6, 2010
Published online: April 29, 2010

Is Longer-Term Psychodynamic Psychotherapy More Effective than Shorter-Term Therapies? Review and Critique of the Evidence

Sunil S. Bhar[a] Brett D. Thombs[b] Monica Pignotti[c] Marielle Bassel[b]
Lisa Jewett[b] James C. Coyne[d] Aaron T. Beck[d]

[a] Swinburne University of Technology, Hawthorn, Vic., Australia; [b] McGill University and Jewish General Hospital, Montréal, Qué., Canada; [c] Florida State University, Tallahassee, Fla., and [d] University of Pennsylvania, Philadelphia, Pa., USA

Key Words
Effectiveness · Efficacy · Long-term therapy · Meta-analysis · Psychotherapy · Systematic review

Abstract
Background: In 2008, Leichsenring and Rabung performed a meta-analysis of 8 studies of longer-term psychodynamic psychotherapy (LTPP). The work was published in the Journal of the American Medical Association (vol. 300, pp 1551–1565), and they concluded that LTPP was more effective than shorter-term therapies. ***Method:*** Given that such claims have the potential to influence treatment decisions and policies, we re-examined the meta-analysis and the 8 studies. ***Results:*** We found a miscalculation of the effect sizes used to make key comparisons. Claims for the effectiveness of LTPP depended on a set of small, underpowered studies that were highly heterogeneous in terms of patients treated, interventions, comparison-control groups, and outcomes. LTPP was compared to 12 types of comparison-controls, including control groups that did not involve any psychotherapy, short-term psychodynamic psychotherapy, and unvalidated treatments. Additionally, the studies failed to protect against threats to bias, and had poor internal validity. ***Conclusion:*** Overall, we found no evidence to support claims of superiority of LTPP over shorter-term methods of psychotherapy.

Introduction

Evidence for the relative effectiveness of psychodynamic treatments is mixed [1]. Flawed study methodologies may at least partially explain inconsistencies in reported findings. The authors of a recent review found 'not a single well-done randomized controlled trial of dynamic therapy' for most mental health disorders [1, p. 103]. Yet, a recent meta-analysis by Leichsenring and Rabung [2] concluded that longer-term psychodynamic psychotherapy (LTPP), defined as lasting at least 1 year or 50 sessions, was 'significantly superior to shorter-term methods of psychotherapy with regard to overall outcome, target problems, and personality functioning' (p. 1563). This claim was based on 8 small studies [3–10][1] that compared LTPP to various control conditions across a diverse range of outcomes.

In an accompanying editorial, Glass [11] declared the meta-analysis to be a 'carefully performed' (p. 1589) defense against criticisms that LTPP lacked empirical support. The meta-analysis has now also been cited as evidence for the effectiveness of psychodynamic treatments [12], and referenced in draft guidelines for empirically supported treatments for depression [13].

[1] As listed on p. 1559 of [2].

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/pps

Sunil S. Bhar
Faculty of Life and Social Sciences, Swinburne University of Technology
Mail H24, PO Box 218
Hawthorn, Vic. 3122 (Australia)
Tel. +61 3 9214 8371, Fax +61 3 9214 5898, E-Mail sbhar@swin.edu.au

An uncritical acceptance of the conclusions of meta-analyses is common [14, 15]. However, as demonstrated by Sensky [16], many meta-analyses are not rigorous and can obscure variations in treatment outcomes across research trials. Failure to adequately specify research questions in systematic reviews can result in 'spurious groupings of data from differing individual studies' [16, p. 132] that do not reflect the efficacy or effectiveness of any specific therapeutic intervention for any particular patient group. Thus, calls have been made for rigorous scrutiny of meta-analyses, including the quality of evidence from individual studies, the degree of clinical heterogeneity in samples, control conditions and outcomes and the methods used to generate aggregate effect sizes [17–20].

Brief post-publication letters [21–24] highlighted serious problems with Leichsenring and Rabung's [2] meta-analysis. In this article, we examine in greater detail the validity of its conclusions. First, we consider the methods employed to aggregate and compare effect sizes across studies. Second, we examine the statistical power and heterogeneity of patients treated, interventions applied and outcomes of reviewed studies. Finally, we assess the risk of bias in the individual studies. Although our critique focuses on one specific meta-analysis, it highlights the types of issues that need to be taken into account when interpreting findings of meta-analyses of comparative outcome studies.

Miscalculation of Effect Sizes

Leichsenring and Rabung [2] calculated separate *within-group* pre-post effect sizes for LTPP and comparison groups. They then erroneously calculated point biserial correlations of group (LTPP vs. comparison) and *within-group* effect sizes. This departed from standard methods where between-group effect sizes are presented as standardized group differences in treatment outcome scores between the 2 groups or as point biserial correlations between group (e.g. LTPP vs. shorter-term therapies) and outcome scores. Standardized mean difference and correlational metrics are essentially equivalent and convertible using simple formulae or tables [25]. Leichsenring and Rabung apparently used a conversion formula intended for conversions of between-group point biserial correlations to standardized difference effect sizes in an attempt to convert their correlations of group and *within-group* pre-post effect sizes into deviation-based effect sizes. As a result, even though none of the 8 studies reported an overall standardized mean difference greater than 1.45 [2, see fig. 2 on p. 1558], the authors reported a combined effect size of 1.8. Similarly, these methods generated an implausible between-group effect size of 6.9, equivalent to 93% of variance explained, for personality functioning based on 4 studies [3, 5, 6, 26][2], none of which reported an effect size more than approximately 2.

Table 1. Pre-post effect size from 10 hypothetical studies

Study	Treatment pre-post effect size	Control pre-post effect size
1	1.00	0.90
2	1.00	0.90
3	1.00	0.90
4	1.00	0.90
5	1.00	0.90
6	1.00	0.90
7	1.00	0.90
8	1.00	0.90
9	1.00	0.90
10	1.00	0.90

Standardized effect size = 46.7 based on methods described by Leichsenring and Rabung [2].

Table 1 shows results from a hypothetical meta-analysis to illustrate the computational error. In 9 of the 10 studies in table 1, the pre-post effect size for the treatment group was 0.10 larger than the effect size for the control group. In the tenth study, the effect size was 0.09 larger. Despite minimal differences in pre-post treatment effects, the method employed by Leichsenring and Rabung [2] generates a correlation between pre-post effect size and group of 0.999 and an implausibly large deviation-based effect size of 46.7. Thus, rather than realistic estimates of the comparative effects of LTPP, Leichsenring and Rabung based their meta-analysis on grossly incorrect calculations.

Problems when Combining Studies

Beyond the issue of statistical aggregation, it is important to ask whether a set of studies can meaningfully be summarized by a single effect estimate [16]. Two factors are important to consider: (1) the extent to which the

[2] Levy et al. [26] analyzed data on personality variables relevant to the outcome study reported by Clarkin et al. [10].

Table 2. Sample size and power of included studies

Study	Treatment n	Control n	Detectable difference (δ) with 70% power	Power to find a moderate (0.50) effect size
Piper et al., 1984 [3]	30	27	0.67	0.46
Bachar et al., 1999 [5]	17	10	1.03	0.23
	17	17	0.88	0.29
Bateman and Fonagy, 1999 [6]	19	19	0.83	0.32
Dare et al., 2001 [7]	21	22	0.78	0.36
	21	19	0.81	0.34
	21	22	0.78	0.36
Svartberg et al., 2004 [8]	25	25	0.72	0.41
Korner et al., 2006 [9]	29	31	0.65	0.48
Clarkin et al., 2007 [10]	30	22	0.71	0.42
	30	17	0.77	0.36
Gregory et al., 2008 [4]	15	15	0.94	0.26

global effect size estimates depend on small, underpowered studies, and (2) the clinical heterogeneity of the patients, interventions, and outcomes.

Power

Leichsenring and Rabung [2] argued that publication bias was absent on the basis of rank correlations between effect size and sample size. However, a non-significant test cannot rule out publication bias, particularly in small meta-analyses with only a few studies included [27]. With only 8 studies being considered, the exercise of calculating such correlations becomes meaningless because non-significant results are virtually foreordained.

Kraemer et al. [28] have further shown that inclusion of small, underpowered trials in meta-analyses results in substantially *overestimated* pooled effect sizes, due to a confirmatory publication bias for which statistical correction is not possible. Kraemer [pers. commun., 2008] proposed that trials included in meta-analyses should have at least a 0.70 probability of detecting a moderate size effect (e.g. $\delta = 0.50$), which would require at least 50 patients per group. The 8 studies pooled by Leichsenring and Rabung [2] had between 15 and 30 patients in the LTPP treatment group, with power to find a moderate effect size ranging between 0.23 and 0.48 (table 2). With the assumption that few investigators would attempt to publish negative studies with such small samples and that few journals would likely accept them for publication, these studies would have had to have an effect size of at least 0.50 to 0.75, the minimum for statistical significance in order to be published and available for review. This guarantees a spurious large effect even when a treatment lacks efficacy.

Heterogeneity

Anticipating the likely *statistical heterogeneity* in the studies being considered, Leichsenring and Rabung [2] used a random effects model for analyses. These models, however, are not capable of compensating for considerable *clinical heterogeneity* in terms of control-comparison groups, diagnoses and outcome measures.

Heterogeneity of Control-Comparison Conditions. The 12 different comparison-control conditions deemed 'shorter-term methods of psychotherapy' included empirically supported treatments, such as dialectic behavior therapy, but also conditions such as waitlist control condition, nutritional counseling, standard psychiatric care, low contact routine treatment, treatment as usual in the community, referrals to alcohol rehabilitation and provision of a therapist phone number (table 3). In one study, the authors explicitly stated that participants in the 'control group received no formal psychotherapy' [6, p. 1565]. To put these control conditions under the umbrella of 'shorter-term methods of psychotherapy' [2, p. 156] is misleading. Even among the studies that compared LTPP to psychotherapy, considerable heterogeneity was evident in the comparison-controls. In only 2 studies was LTPP compared to an *empirically supported treatment*, as defined by Chambless and Hollon [29], that is, dialectic behavior therapy for borderline personality disorder [10], and family therapy for anorexia nervosa [7]. In 2 other studies, LTPP was compared to cognitive therapy [8] and short-term psychodynamic psychotherapy [3], which are *established* as efficacious for some disorders, but not yet validated for the disorder being treated (i.e. cluster C personality disorders, 'neurosis'). In a fifth study [5], LTPP was compared to 'cognitive orientation therapy', an unvalidated treatment. In these original studies, statistical superiority of LTPP over control conditions was found only when control conditions involved either no psychotherapy, or an unvalidated treatment. Studies that compared LTPP to an empirically supported (e.g. dialectic behavior therapy, family therapy) or established treatment (e.g. short-term psychodynamic psychotherapy, cognitive therapy) found that LTPP was equally or less effective than these treatments despite a substantially longer treatment period.

Heterogeneity of Diagnoses. The samples in these studies were heterogeneous in terms of patient diagnoses.

Table 3. Characteristics of control interventions classified as 'shorter-term methods of psychotherapy' in included studies

Control intervention	Therapy	EST[1]	Duration months	Mental disorder	Statistically superior treatment	Primary source
STPP (Malan)	yes	no	6	neurosis, PD, no diagnosis	STPP	Piper et al., 1984 [3]
Nutritional counsel	no	no	6	anorexia nervosa and Bulimia	LTPP	Bachar et al., 1999 [5]
Cognitive orientation therapy	yes	no	12	anorexia nervosa and Bulimia	LTPP	Bachar et al., 1999 [5]
TAU – psychiatric care	no	no	18	BPD	LTPP	Bateman and Fonagy, 1999 [6]
Cognitive-analytic	yes	no	7	anorexia nervosa	same	Dare et al., 2001 [7]
Family therapy	yes	yes	12	anorexia nervosa	same	Dare et al., 2001 [7]
TAU – routine treatment	no	no	12	anorexia nervosa	LTPP	Dare et al., 2001 [7]
CT	yes	no	10	cluster C or self-defeating PD	same	Svartberg et al., 2004 [8]
TAU – waitlist	no	no	12	BPD	LTPP	Korner et al., 2006 [9]
Dynamic supportive	yes	no	12	BPD	same	Clarkin et al., 2007 [10]
DBT	yes	yes	12	BPD	same	Clarkin et al., 2007 [10]
TAU – community	no	no	12	BPD with co-morbid AD	LTPP	Gregory et al., 2008 [4]

AD = Alcohol use disorder; BPD = borderline personality disorder; DBT = dialectic behavior therapy; EST = empirically supported treatment; LTPP = longer-term psychodynamic psychotherapy; PD = personality disorder; STPP = short-term psychoanalytic psychotherapy; TAU = treatment as usual.

[1] Treatments were classified as EST if they met criteria for being a 'well established treatment' or 'probable efficacious treatment' as defined by Chambless and Hollon [29] and listed as such by the American Psychological Society, Division 12 Society of Clinical Psychology on their website, www.psychology.sunysb.edu/eklonsky-/division12/index.html.

Across the 8 studies, LTPP was compared to other interventions for a total of 9 types of mental health problems, including 'neurosis' [3], 'self-defeating personality disorder' [8] and anorexia nervosa [5, 7] (table 3). This is akin to asking whether one type of medication is superior to another for all types of physical illnesses [30]. As recommended by Chambless and Hollon [29], treatment outcome research should 'not ask whether a treatment is efficacious; rather […] whether it is efficacious for a specific problem or population' (p. 9). A clinician could not determine from the global effect sizes published by Leichsenring and Rabung [2] whether to expect a patient with an eating disorder or with Borderline Personality Disorder, for instance, to benefit from LTPP. For that information, a clinician would have to return to the original studies, which were too small and methodologically weak to offer much guidance.

Heterogeneity of Outcomes. The number of outcomes in the 8 studies ranged from two [9] to twelve [3]. The standard practice of selecting a single primary outcome measure may be difficult to accomplish in complex interventions, such as psychotherapy [31]. Nonetheless, combining all outcomes reported in a given dimension and then considering each dimension equally in producing a single study effect size was poorly justified and resulted in questionable estimates of effect sizes. Specifically, Leichsenring and Rabung [2] classified each measure as an indicator of 1 of 4 dimensions: *target problems*, *symptoms*, *social functioning*, or *personality*. They then averaged the effect size of all measures within each dimension to obtain a dimension effect size. From there, they averaged dimension effect sizes to obtain a single global effect size for a study. One study [5], for example, reported 2 outcome measures of *target problems,* one with an extremely large effect size of 1.46, and the other with a negligible effect size of 0.17. These were apparently averaged to obtain a single large effect size of 0.82. This effect size of 0.82 for *target problems* was then averaged with an effect size of essentially zero (0.01) for the *symptoms* dimension, and a very large effect size of 1.13 for the *personality dimension* to produce a global effect size of 0.65 that was not representative of any of the individual effect size estimates for any of the outcomes.

Assessment of Risk of Bias in the Included Studies

The validity of a meta-analysis is limited by the quality of evidence provided by the individual studies. As stated by Chambless and Hollon [29], it is 'unwise to rely on

Table 4. Assessment of risk of bias in studies on the comparative effects of LTPP to shorter-term methods of psychotherapy

Study	Adequate sequence generation	Allocation concealment	Blinding of assessors	All outcome data addressed	Free of selective reporting	Controlled for frequency of sessions per week	Controlled for the imbalance in treatment augmentation
1. Piper et al., 1984 [3]	no	yes	unclear	no	unclear	yes	no
2. Bachar et al., 1999 [5]	unclear	unclear	unclear	no	unclear	no	unclear
3. Bateman and Fonagy, 1999 [6]	unclear	unclear	unclear	unclear	unclear	no	yes
4. Dare et al., 2001 [7]	yes	unclear	no	unclear	unclear	no	unclear
5. Svartberg et al., 2004 [8]	unclear	unclear	yes	unclear	unclear	yes	yes
6. Korner et al., 2006 [9]	no	no	no	unclear	unclear	no	unclear
7. Clarkin et al., 2007 [10]	unclear	unclear	yes	yes	yes	no	yes
8. Gregory et al., 2008 [4]	yes	unclear	no	unclear	yes	no	no

No = High risk of bias; unclear = unclear risk of bias; yes = low risk of bias. A more detailed report on coding rationale can be obtained from the investigators.

meta-analyses unless something is known about the quality of studies that have been included and there is confidence in the data.' (p. 13). Leichsenring and Rabung [2] assessed the quality of studies with a modified version of the Jadad scale [32] and found no significant association between the total score on this 3-item scale and post-treatment effect sizes. They concluded that study quality was not related to effect sizes.

This approach was problematic for 3 reasons. First, single quality scores are known to be unreliable and their use is not supported by empirical evidence. Second, the Jadad scale is not designed as a generic measurement of study quality, but rather emphasizes completeness of reporting, and therefore would not be expected to discriminate between studies or identify poorly designed studies. For example, it does not cover one of the most important potential biases in randomized trials, namely allocation concealment [18]. Third, relying on correlations between the total score of such scales and effect sizes is not informative about the overall quality of evidence proffered by the sample of studies [18]; the lack of association may reflect problems such as poor statistical power or restricted range when many poor-quality studies are combined. Rather than the methods used by Leichsenring and Rabung, the quality of evidence is best assessed by examining whether comprehensive procedures were in place to protect against specific sources of bias [18].

The Cochrane Collaboration's tool [18] addresses 6 sources of bias: methods of sequence generation; procedures for concealing allocations of participants to conditions; whether blinding of assessors or participants was established; completeness in outcome data; omissions in reporting outcomes, and other sources of potential bias (e.g. imbalance between interventions, poor treatment integrity). Following the Cochrane guidelines, we evaluated the risk of bias in each of the 8 studies sampled by Leichsenring and Rabung [2] and applied a rating of 'Yes', 'No' or 'Unclear' to denote whether adequate measures were taken to protect against each potential source of bias in each study, with 'Yes' indicating that such measures were taken, 'No' that they was not, and 'Unclear' indicating insufficient information. Two investigators independently assessed each article and inconsistencies were resolved by consensus. Inter-rater agreement was high for pre-consensus ratings. Of the 56 ratings shown in table 4 (assessment of risk of bias, except for treatment integrity), 52 (94.6%) were matched between raters (Cohen's kappa = 0.92). Of the 64 ratings of treatment integrity procedures in table 5 (assessment of treatment integrity), 58 (90.6%) were initially matched between raters (Cohen's kappa = 0.80).

Adequate Sequence Generation

As shown in table 4, only 2 of the 8 studies identified methods of randomization [4, 7]. Two studies clearly did not randomize patients to conditions. In one, patients were not randomly assigned across long- and short-term forms of treatment, but only between individual and group-based versions of the treatment [3]. The other included a non-random waitlist control group [9]. The remaining 4 studies [5, 6, 8, 10] stated that patients were randomly assigned, but did not provide information to determine how this was done.

Table 5. Evaluation of whether various treatment integrity procedures were described in the included studies

Study	Established integrity procedures		Assessed integrity		Evaluated threats to integrity data		Reported numerical information	
	used manual	supervised therapists	adherence assessed	competence assessed	controlled for therapist reactivity	evaluated interrater reliability	adherence data	competence data
1. Piper et al., 1984 [3]	no	unclear	no	no	no	no	no	no
2. Bachar et al., 1999 [5]	yes	yes	yes	unclear	no	no	no	no
3. Bateman and Fonagy, 1999 [6]	yes	yes	yes	no	no	no	no	no
4. Dare et al., 2001 [7]	yes	yes	no	no	no	no	no	no
5. Svartberg et al., 2004 [8]	yes	yes	yes	yes	no	no	no	no
6. Korner et al., 2006 [9]	yes	yes	no	no	no	no	no	no
7. Clarkin et al., 2007 [10]	yes	yes	yes	yes	no	no	no	no
8. Gregory et al., 2008 [4]	yes	yes	yes	yes	no	no	no	no

No = High risk of bias; unclear = unclear risk of bias; yes = low risk of bias.

Concealment of Allocation Sequence

Adequate concealment of allocation sequence shields those who enroll participants from knowing upcoming assignments. Such knowledge could allow selective enrollment of participants to favor a particular treatment. Piper et al. [3] successfully concealed the allocation sequences by delaying decisions about assignments until after initial assessments. However, the remaining 7 studies either failed to conceal allocation sequence [9] or did not provide sufficient information to assess allocation concealment [4–8, 10]. For example, Gregory et al. [4] reported that patients were randomly assigned to outpatient treatment conditions, but did not provide information about measures taken to prevent investigators knowing ahead of time what the assignments would be.

Blinding

Blinding refers to a process by which study participants and personnel, including raters assessing outcomes, are kept unaware of intervention allocations after enrollment of participants [18]. In psychotherapy research, complete blinding of participants and providers is not possible. However, independent assessment of outcomes by raters blinded to the treatment received can reduce bias.

Blinding was not judged to be a source of bias in 2 studies [8, 10] where outcomes were measured through self-report measures. In 3 studies, information was not provided on whether or not outcome assessment was blind [3, 5, 6]. For example, Bachar et al. [5] reported that pre- and post-assessment procedures were conducted by the same evaluator, who was not the therapist, but did not specify if the evaluator was blinded to group assignment. In the remaining 3 studies, blindness of outcome evaluators was either not ensured [7, 9] or not adequately tested [4]. Guessing group assignment [4] is an unsatisfactory test of blindness [33].

Incomplete Outcome Data

Missing outcome data due to drop-outs during treatment or exclusion of enrolled participants from analyses can generate substantially biased effect estimates [18]. Intention-to-treat (ITT) analysis is the recommended standard by which outcome effects should be evaluated in order to avoid this problem [34]. An ITT analysis is conducted on all randomized participants, regardless of whether they completed the intervention, and incomplete outcome data are replaced using imputation strategies. The imputation strategy of last observation carried forward (LOCF) is unsatisfactory because it can artificially inflate or deflate the difference between experimental and treatment groups [35, 36].

In table 4, studies were rated relatively low in risk of bias if they employed an ITT analysis and described an optimal strategy for imputing missing data (e.g. growth curve analysis). Studies were rated as having high risk of bias, if they analyzed only those individuals who completed interventions (i.e. 'completers' analysis). Studies that used LOCF for imputing missing data or did not describe their imputation strategy were rated as having an unclear level of risk. Only 1 of the 8 studies met criteria for low risk of bias [10]. Two studies were classified as hav-

ing a relatively high risk of bias, because they analyzed completers only [3, 5]. The remaining 5 studies were rated as having an unclear level of risk because they used variants of LOCF to estimate missing data in their ITT analysis [4, 7], did not report whether data were missing [9], or did not describe if and how missing data were estimated [6, 8].

Selective Outcome Reporting

Selective outcome reporting refers to reporting a subset of the outcomes, based on the results obtained [18], such as only reporting statistically significant results. We checked if studies reviewed by Leichsenring and Rabung [2] were registered (clinicaltrials.gov database, World Health Organization International Clinical Trials Registry Platform). Of the 8 studies we reviewed, the 3 most recent studies could potentially have been registered prior to publication [4, 9, 10], but only 1 was registered and pre-specified its outcome variables [4]. For the remaining 7 studies, we examined the article for an explicit statement that a priori determined outcomes were reported. One study contained such a statement [10]. For the other 6 studies, the extent to which outcomes were selectively reported remained unclear (table 4).

Other Potential Threats to Validity

We next considered the degree to which the included studies protected against bias due to *treatment imbalance* and *treatment integrity*. *Treatment imbalance* refers to uncontrolled differences between how treatments were administered that may have accounted for why 1 treatment was superior to another. Three types of differences were identified between treatment groups: (1) number of sessions and duration of treatment; (2) frequency of sessions per week, and (3) medication augmentation. First, patients in the LTPP condition received on average, 103 sessions over 53 weeks, while those in the comparison conditions received on average, 33 sessions over 39 weeks [2]. Second, 6 of the 8 studies did not control differences in session frequency between conditions. In one study [7], LTPP consisted of weekly sessions, while family therapy was 'scheduled by negotiation between once a week and once every 3 weeks' (p. 217). In another study, patients in LTPP received twice as many sessions per week, compared to those in supportive treatment [10]. Third, confounding concurrent medications were not controlled in 5 of the 8 studies [3–5, 7, 9]. Such omissions preclude evaluation of the independent effect of LTPP.

Treatment integrity, which refers to whether treatments were implemented as intended and whether conclusions about relative efficacy may reflect differential competency in delivery [37, 38], was not adequately examined in any of the 8 studies. Assessments of treatment integrity should reflect both whether therapists adhered to treatment protocols (treatment adherence) and the degree to which they delivered treatments competently. Treatment adherence and competence can each be evaluated on 4 domains reflecting how well each aspect was: (1) established (e.g. did therapists use treatment manuals and were they trained to competently administer the treatments?); (2) assessed (e.g. did researchers use sound psychometric measures to assess adherence and competence?); (3) evaluated (e.g. were procedures adopted to minimize therapist reactivity, that is, to minimize the extent to which therapists modify behavior when being observed; were raters reliable?), and (4) reported (e.g. did researchers provide numerical descriptions of levels of adherence or competence?) [38].

The use of traditional treatment manuals has typically been eschewed by proponents of psychoanalytic treatment [39] who argue that such manuals constrict that therapeutic approach and thus reduce the effectiveness of the treatment [40]. Nevertheless, 7 of the 8 studies reviewed by Leichsenring and Rabung [2] employed a manual or manual-like protocol to guide treatment (table 5). However, other gaps in the examination of treatment integrity were found more frequently among the sample of studies. Three studies did not assess treatment adherence [3, 7, 9], and 4 did not describe procedures to assess therapist competency [3, 6, 7, 9]. None controlled for therapist reactivity, evaluated inter-rater variability, or reported numerical ratings of levels of adherence or competence. Overall, the extent to which treatments were delivered as intended in this sample of studies was poorly documented.

Conclusions

Our examination of Leichsenring and Rabung's [2] meta-analysis demonstrates how the failure to attend to important methodological issues can generate invalid conclusions. The most damning threat to the validity of findings in their meta-analysis was their gross miscalculation and aggregation of comparative effect sizes of included studies. Second, the meta-analysis was based on studies with small samples sizes, and overly heterogeneous comparative treatments, disorders and outcomes. Thus, one cannot know if LTPP is superior to other treatments only for certain disorders, but not for others, or for

none. Given that treatment integrity was not well examined in most studies, one also cannot know if the delivery of treatments was executed in a competent manner and in the manner described in study treatment protocols. Finally, their conclusion was based on a sample of studies that failed to protect against multiple potential sources of bias, thus producing weak quality evidence.

Our assessment is consistent with Gibbons et al. [1] who found the literature concerning dynamic therapy to be methodologically weak. Our observations also mirror other commentaries regarding the methodological limitations of meta-analyses [14, 16]. Despite the recognition of the importance of strategies to limit bias in the assembly and analysis of primary studies included in meta-analyses [18], our review of Leichsenring and Rabung's [2] study shows that meta-analytic techniques continue to be poorly implemented even in high-impact journals.

There are a number of caveats to our assessment. First, we assessed the risk of bias, rather than the true level of bias in included studies. Second, our assessment of the articles was based on what was reported rather than what was done in the research. Authors of articles were not contacted for information. It is possible that methodological procedures were omitted from the reports or were misrepresented. Third, although our critique covered an extensive range of potential methodological flaws affecting the meta-analysis, other factors were not addressed in this review, such as the reliability of measurements, criteria employed for selecting a study for inclusion into the meta-analysis, and effects of excluding relevant articles. In addition, we did not employ assessment tools such as AMSTAR [41] to formally evaluate the reporting or methodological quality of the meta-analysis. Therefore, the quality of the meta-analysis as defined by these tools is yet to be investigated.

Our critique highlights the type of problems that can be inherent in meta-analyses and which need to be addressed by researchers employing this research strategy. As meta-analyses become more prominent in the field, specific attention should also be directed towards the quality of evidence, synthesis of the evidence, and statistical methods of aggregating effect sizes. As shown in our review, the failure to adequately attend to each of these potential problems reduces the scientific credibility of meta-analyses. Thus, the question of whether LTPP is more effective than shorter-term psychotherapy remains open. Answering that question would require more rigorous comparative trials than are currently available.

Acknowledgements

This research was supported by the National Institute of Mental Health (P30 MH45178). Dr. Thombs is supported by a New Investigator Award from the Canadian Institutes of Health Research and an Établissement de Jeunes Chercheurs award from the Fonds de la Recherche en Santé Québec. Ms. Jewett is supported by a Frederick Banting and Charles Best Canadian Graduate Scholarship – Master's Award from the Canadian Institutes of Health Research. We wish to thank Gloria Huh and Megan Spokas for their assistance in revising the manuscript.

References

1 Gibbons MBC, Crits-Christoph P, Hearon B: The empirical status of psychodynamic therapies. Annu Rev Clin Psychol 2008;4:93–108.
2 Leichsenring F, Rabung S: Effectiveness of long-term psychodynamic psychotherapy: a meta-analysis. JAMA 2008;300:1551–1565.
3 Piper WE, Debbane EG, Bienvenu JP, Garant J: A comparative study of four forms of psychotherapy. J Consult Clin Psychol 1984;52:268–279.
4 Gregory RJ, Chlebowski S, Kang D, Remen AL, Soderberg MG, Stepkovitch J, Virk S: A controlled trial of psychodynamic psychotherapy for co-occurring borderline personality disorder and alcohol use disorder. Psychother Theor Res Pract Train 2008;45:28–41.
5 Bachar E, Latzer Y, Kreitler S, Berry EM: Empirical comparison of two psychological therapies: self psychology and cognitive orientation in the treatment of anorexia and bulimia. J Psychother Pract Res 1999;8:115–128.
6 Bateman AW, Fonagy P: Effectiveness of partial hospitalization in the treatment of borderline personality disorder: a randomized controlled trial. Am J Psychiatry 1999;156:1563–1569.
7 Dare C, Eisler I, Russell G, Treasure J, Dodge L: Psychological therapies for adults with anorexia nervosa: randomised controlled trial of out-patient treatments. Br J Psychiatry 2001;178:216–221.
8 Svartberg M, Stiles TC, Seltzer MH: Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for cluster c personality disorders. Am J Psychiatry 2004;161:810–817.
9 Korner A, Gerull F, Meares R, Stevenson J: Borderline personality disorder treated with the conversational model: a replication study. Compr Psychiatry 2006;47:406–411.
10 Clarkin JF, Levy KN, Lenzenweger MF, Kernberg OF: Evaluating three treatments for borderline personality disorder: a multiwave study. Am J Psychiatry 2007;164:922–928.
11 Glass RM: Psychodynamic psychotherapy and research evidence: Bambi survives Godzilla? JAMA 2008;300:1587–1589.
12 de Maat S, de Jonghe F, Schoevers R, Dekker J: The effectiveness of long-term psychoanalytic therapy: a systematic review of empirical studies. Harv Rev Psychiatry 2009;17:1–23.
13 National Institute for Clinical Excellence: Depression: the treatment and management of depression in adults. National clinical practice guideline number X. London, NICE, 2009.
14 Delaney A, Bagshaw SM, Ferland A, Manns B, Laupland KB, Doig CJ: A systematic evaluation of the quality of meta-analyses in the

critical care literature. Crit Care 2005; 9:R575–R582.
15 Dixon E, Hameed M, Sutherland F, Cook DJ, Doig C: Evaluating meta-analyses in the general surgical literature: a critical appraisal. Ann Surg 2005;241:450–459.
16 Sensky T: The effectiveness of cognitive therapy for schizophrenia: what can we learn from the meta-analyses? Psychother Psychosom 2005;74:131–135.
17 Lipsey MW, Wilson DB: Practical Meta-Analysis. London, Thousand Oaks, 2001.
18 Higgins JP, Altman DG: Assessing risk of bias in included studies; in Higgins J, Green S (eds): Cochrane collaboration for systematic reviews of interventions. The Cochrane Collaboration, 2008. www.cochrane-handbook.org
19 Egger M, Smith GD, Sterne JAC: Uses and abuses of meta-analysis. Clin Med 2001;1: 478–484.
20 Moher D, Schulz KF, Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T: The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001; 134:657–662.
21 Roepke S, Renneberg B: Analyzing effectiveness of long-term psychodynamic psychotherapy. JAMA 2009;301:931–932.
22 Thombs BD, Bassel M, Jewett L: Analyzing effectiveness of long-term psychodynamic psychotherapy. JAMA 2009;301:930.
23 Kriston L, Holzel L, Harter M: Analyzing effectiveness of long-term psychodynamic psychotherapy. JAMA 2009;301:930–931.
24 Beck AT, Bhar SS: Analyzing effectiveness of long-term psychodynamic psychotherapy. JAMA 2009;301:931.
25 McCartney K, Rosenthal R: Effect size, practical importance, and social policy for children. Child Dev 2000;71:173–180.
26 Levy KN, Meehan KB, Kelly KM, Reynoso JS, Weber M, Clarkin JF, Kernberg OF: Change in attachment patterns and reflective function in a randomized control trial of transference-focused psychotherapy for borderline personality disorder. J Consult Clin Psychol 2006;74:1027–1040.
27 Begg CB, Mazumdar M: Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088–1101.
28 Kraemer HC, Gardner C, Brooks JO III, Yesavage JA: Advantages of excluding underpowered studies in meta-analysis: inclusionist versus exclusionist viewpoints. Psychol Methods 1998;3:23–31.
29 Chambless DL, Hollon SD: Defining empirically supported therapies. J Consult Clin Psychol 1998;66:7–18.
30 DeRubeis RJ, Brotman MA, Gibbons CJ: A conceptual and methodological analysis of the nonspecifics argument. Clin Psychol Sci Pract 2005;12:174–183.
31 Scott J, Sensky T: Methodological aspects of randomized controlled trials of psychotherapy in primary care. Psychol Med 2003;33: 191–196.
32 Jadad AR, Moore RA, Carrol D, Jenkinson C, Reynolds DJM, Gavaghan DJ, McQuay HJ: Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1–12.
33 Sharpe L, Ryan B, Allard S, Sensky T: Testing for the integrity of blinding in clinical trials: how valid are forced choice paradigms? Psychother Psychosom 2003;72:128–131.
34 Newell DJ: Intention-to-treat-analysis: implications for quantitative and qualitative research. Int J Epidemiol 1992;21:837–841.
35 Streiner DL: The case of the missing data: methods of dealing with dropouts and other research vagaries. Can J Psychiatry 2002;47: 68–75.
36 Sensky T: Evaluating the evidence base of consultation-liaison psychiatry; in Streltzer J, Hoyle L (eds): Handbook of Consultation-Liaison Psychiatry. New York, Springer, 2008, pp 28–36.
37 Waltz J, Addis ME, Koerner K, Jacobson NS: Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. J Consult Clin Psychol 1993;61:620–630.
38 Perepletchikova F, Treat TA, Kazdin AE: Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. J Consult Clin Psychol 2007;75:829–841.
39 Fredric B, Barbara M, Larry S: A study demonstrating efficacy of a psychoanalytic psychotherapy for panic disorder: implications for psychoanalytic research, theory, and practice. J Am Psychoanal Assoc 2009;57: 131–148.
40 Blatt SJ: The effort to identify empirically supported psychological treatments and its implications for clinical research, practice, and training: commentary on papers by Lester Luborsky and Hans H. Strupp. Psychoanal Dialogues 2001;11:635–646.
41 Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM: Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10–17.

Erratum

In the paper by Bhar et al. (Psychother Psychosom 2010;79: 208-216, published online April 29, 2010), there was an error in table 1. The corrected version of table 1 should read:

Table 1. Pre-post effect size from 10 hypothetical studies

Study	Treatment pre-post effect size	Control pre-post effect size
1	1.00	0.90
2	1.00	0.90
3	1.00	0.90
4	1.00	0.90
5	1.00	0.90
6	1.00	0.90
7	1.00	0.90
8	1.00	0.90
9	1.00	0.90
10	1.00	0.91

Standardized effect size = 46.7 based on methods described by Leichsenring and Rabung [2].