A Series of Meta-Analytic Tests of the Efficacy of Long-Term Psychoanalytic Psychotherapy

Christian Franz Josef Woll[1] and Felix D. Schönbrodt[2]

[1]Department of Psychology, Clinical Psychology of Children and Adolescents and Psychology of Interventions, Ludwig-Maximilians-Universität Munich, Germany
[2]Department of Psychology, Psychological Methods and Assessment, Ludwig-Maximilians-Universität Munich, Germany

Abstract: Recent meta-analyses come to conflicting conclusions about the efficacy of long-term psychoanalytic psychotherapy (LTPP). Our first goal was to reproduce the most recent meta-analysis by Leichsenring, Abbass, Luyten, Hilsenroth, and Rabung (2013) who found evidence for the efficacy of LTPP in the treatment of complex mental disorders. Our replicated effect sizes were in general slightly smaller. Second, we conducted an updated meta-analysis of randomized controlled trials comparing LTPP (lasting for at least 1 year and 40 sessions) to other forms of psychotherapy in the treatment of complex mental disorders. We focused on a transparent research process according to open science standards and applied a series of elaborated meta-analytic procedures to test and control for publication bias. Our updated meta-analysis comprising 191 effect sizes from 14 eligible studies revealed small, statistically significant effect sizes at post-treatment for the outcome domains psychiatric symptoms, target problems, social functioning, and overall effectiveness (Hedges' *g* ranging between 0.24 and 0.35). The effect size for the domain personality functioning (0.24) was not significant (p = .08). No signs for publication bias could be detected. In light of a heterogeneous study set and some methodological shortcomings in the primary studies, these results should be interpreted cautiously. In conclusion, LTPP might be superior to other forms of psychotherapy in the treatment of complex mental disorders. Notably, our effect sizes represent the additional gain of LTPP versus other forms of primarily long-term psychotherapy. In this case, large differences in effect sizes are not to be expected.

Keywords: efficacy, long-term, psychoanalytic psychotherapy, meta-analysis, publication bias

Since its origin at the end of the 19th century, psychoanalysis has always been a controversial issue. Some have criticized its scientific standing during the 20th century (Grünbaum, 1988; Popper, 1972), whereas others have corroborated psychoanalytic concepts with empirical evidence (Masling, 1983; Werner & Langenmayr, 2006). To grasp the complexity of psychoanalysis, one first needs to distinguish between psychoanalysis as a theory of the human mind, as a methodology to study human and social phenomena, and as a therapeutic method to treat mental disorders (Mertens, 2013). In our paper, we solely focus on psychoanalysis as a therapeutic method.

In the introduction of their systematic review on long-term psychoanalytic therapy, de Maat, de Jonghe, Schoevers, and Dekker (2009) list 15 studies dated from 1917 till 1968 showing that efforts have been made from the very beginning to prove the effectiveness of psychoanalytic therapy. Recent meta-analyses come to conflicting conclusions about the efficacy of long-term psychoanalytic psychotherapy (Leichsenring, Abbass, Luyten, Hilsenroth, & Rabung, 2013; Smit et al., 2012). In our study, we aim to address this conflicting evidence by replicating the most recent meta-analysis by Leichsenring et al. (2013) and by conducting an updated meta-analysis of long-term psychoanalytic psychotherapy. However, before opening the discourse in more detail, we first outline some basic definitions and the current status of psychoanalytic psychotherapy research.

Long-term psychoanalytic therapy encompasses *long-term psychoanalytic psychotherapy (LTPP)* and *psychoanalysis proper* (de Maat et al., 2013). *Short-term psychoanalytic psychotherapy (STPP)* is a less time requiring and more focused treatment modality. Figure 1 gives a schematic overview of the psychoanalytic treatment modalities, also showing the equivalent use of the terms *psychoanalytic* and *psychodynamic*. There is no generally accepted definition of the dosage of short-term and long-term psychoanalytic psychotherapy. Abbass et al. (2014) define short-term as lasting 40 or fewer sessions in total. However, some researchers begin to divide the range up to 40 sessions into short-term and medium-term (Leichsenring, Leweke, Klein, & Steinert, 2015), which might be useful for an increasing

https://doi.org/10.1027/1016-9040/a000385

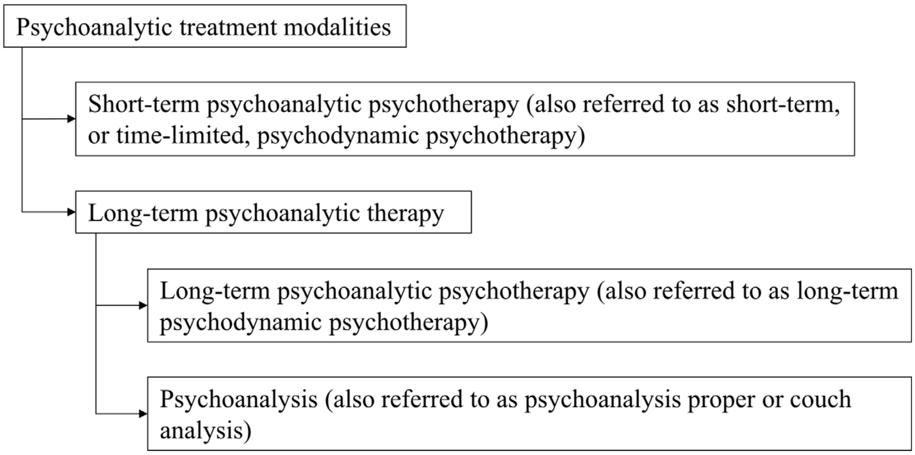

Psychoanalytic treatment modalities
Short-term psychoanalytic psychotherapy (also referred to as short-term, or time-limited, psychodynamic psychotherapy)
Long-term psychoanalytic therapy
Long-term psychoanalytic psychotherapy (also referred to as long-term psychodynamic psychotherapy)
Psychoanalysis (also referred to as psychoanalysis proper or couch analysis)

Figure 1. Psychoanalytic treatment modalities (de Maat et al., 2013, p. 108). Figure available at https://osf.io/vec5d/, under a CC-BY 4.0 license.

data base (Leichsenring & Rabung, 2011). Gabbard (2017) argues that therapies lasting for 40 or more sessions should be regarded as long-term. The difference between the two long-term modalities is most frequently explained by the therapeutic setting (de Maat et al., 2013). In psychoanalysis proper, the therapist is sitting behind the patient lying on a couch, whereas in LTPP, both of them are sitting on chairs facing each other. The frequency of sessions ranges from two to five sessions per week in psychoanalysis proper and from one to two sessions per week in LTPP.

The controversial debate about the efficacy of STPP seems settled since there is a wide range of studies fulfilling the criteria of evidence-based psychotherapy (Abbass, Hancock, Henderson, & Kisely, 2006; Abbass et al., 2014; Leichsenring et al., 2015; Leichsenring, Rabung, & Leibing, 2004). However, every form of psychotherapy has to be rescrutinized as methods progress since effect sizes might be overestimated (Cuijpers, Karyotaki, Reijnders, & Ebert, 2019). Nevertheless, one should carefully consider the choice of the control condition when interpreting an effect size (Munder et al., 2019).

For the two long-term treatment modalities LTPP and psychoanalysis proper, the controversial debate is still going on. To begin with, it is worth noting that many patients, in particular those with chronic mental disorders or personality disorders, may need more extended treatments because shorter forms of psychotherapy are not sufficient enough to cure them effectively and sustainably (Kopta, Howard, Lowry, & Beutler, 1994; Zimmermann et al., 2015). Further findings support this assumption because psychological treatments for mood disorders in adults which offer an additional maintenance and continuation treatment have been found to reduce the risk of recurrence (Hollon & Ponniah, 2010). These findings are consistent with results from meta-analyses suggesting that LTPP is superior to shorter forms of psychotherapy in the treatment of *complex* mental disorders, meaning multiple mental disorders, chronic mental disorders, or personality disorders (Leichsenring et al., 2013; Leichsenring & Rabung, 2011). These meta-analyses only included randomized controlled trials (RCTs) and showed positive results for the efficacy of LTPP. The number of RCTs with 10 and 12 studies in the meta-analyses from 2011 and 2013, respectively, is relatively small, though. Based on an overlapping dataset, Smit et al. (2012) conducted another meta-analysis of RCTs coming to the conclusion that the evidence for an effect of LTPP is limited and conflicting. However, being more in line with the psychoanalytic research tradition, there is a higher number of naturalistic, uncontrolled studies revealing moderate to large effects for LTPP as shown in a systematic review by de Maat et al. (2009).

Considering de Maat et al.'s (2013) meta-analysis of outcome research on psychoanalysis proper, the empirical evidence for this long-term treatment modality is based on a limited number of mainly naturalistic, uncontrolled studies providing evidence for pre to post changes in psychoanalysis patients with complex mental disorders. Only one of the included studies was an RCT, conducted by Huber, Zimmermann, Henrich, and Klug (2012). Currently, Beutel et al. (2016) and Benecke, Huber, Staats, et al. (2016) each run an RCT comparing long-term cognitive behavioral therapy to long-term psychoanalytic therapy (long-term CBT vs. LTPP *and* psychoanalysis proper). Their study designs follow the highest standards of current psychotherapy research. They will soon provide further information on the outcomes of long-term psychoanalytic therapy.

To summarize, at this stage there is considerable evidence for an effect of STPP and, at the same time, the number of RCTs of psychoanalysis proper is too sparse to conduct a reasonable meta-analysis. It was, therefore, our goal to make a contribution towards resolving the conflicting evidence for LTPP (Leichsenring et al., 2013; Smit et al., 2012) by conducting a meta-analysis of RCTs. Focusing on RCTs is required by international guidelines such as the current, updated concept of empirically supported treatments

(ESTs) by the American Psychological Association (APA; Tolin, McKay, Forman, Klonsky, & Thombs, 2015). Notably, even though the concept was updated, it can still be criticized as being too far from clinical, psychotherapeutic reality (e.g., neglecting comorbidity and applying a random assignment; Orlinsky, 2008, Seligman, 1995).

When conducting a meta-analysis, errors are common and the number of effect sizes which cannot be reproduced is much higher than is desirable (Lakens, Hilgard, & Staaks, 2016). Therefore, transparent, reproducible replications and updates of meta-analyses are needed to increase the credibility of meta-analytic conclusions. Another important issue of combining information across multiple studies constitutes the adjustment for *publication bias*, meaning statistically nonsignificant, counterfactual results are less likely to be published (McShane, Böckenholt, & Hansen, 2016). Fanelli (2012) provides overwhelming evidence for publication bias across disciplines and countries. Hence, it is crucial to employ statistical techniques to assess and adjust for publication bias.

In the update of their meta-analysis from 2011, Leichsenring et al. (2013) critically reviewed the meta-analysis conducted by Smit et al. (2012) who doubt an effect of LTPP. Leichsenring et al. (2013) found several methodological shortcomings such as the inclusion of studies that regarding content do not represent LTPP. Additionally, Smit et al. (2012) included two studies that do not meet their own inclusion criterion of the length of LTPP. Furthermore, Smit et al. (2012) actually compared LTPP to other forms of long-term therapy as indicated by a mean session ratio of 1.35 (i.e., mean number of sessions in LTPP vs. mean number of sessions in the control group). They found LTPP to be as efficacious and, thus, did not really find results contradicting those of the previous meta-analysis by Leichsenring and Rabung (2011) in which the session ratio was 1.96. Their session ratio of 1.96 indicated that about twice as many sessions were carried out in the LTPP condition as compared to the controls, meaning that they compared LTPP to shorter or less intensive forms of psychotherapy.

Since the meta-analysis by Smit et al. (2012) has already been critically reviewed, we decided to review and update Leichsenring et al.'s (2013) most recent meta-analysis of LTPP. In a first step, we tried to reproduce Leichsenring et al.'s (2013) effect sizes using the same set of studies they had used. Second, we conducted our own meta-analysis (a) updating the data base with recent RCTs of complex mental disorders, (b) following Lakens et al.'s (2016) recommendations for a transparent research process, (c) relating to international guidelines to assess the risk of bias in the primary studies (Higgins et al., 2011), and (d) employing elaborated statistical techniques to assess and adjust for publication bias.

With LTPP causing higher direct financial costs because of a higher number of sessions, its effects need to exceed those of less intensive treatments. We, therefore, compared LTPP to other forms of psychotherapy by regarding the session ratio of each study as a possible continuous moderating variable, implying that a higher session ratio should result in a larger effect size and vice versa. After accounting for publication bias in our meta-analysis, we expected our effect sizes to be smaller than those of Leichsenring et al. (2013) but different from a null effect considering moderate to large effects from naturalistic, uncontrolled studies (de Maat et al., 2009).

Method

For ensuring a reproducible research process, we (a) disclosed all of our meta-analytic data and statistical scripts on the Open Science Framework (see https://osf.io/vec5d/), (b) specified which effect size calculations were used and which assumptions were made for missing data to facilitate quality control, (c) adhered to reporting guidelines, in our case the Meta-Analysis Reporting Standards (MARS; APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) to guarantee a clear and thorough description, (d) pre-registered our introduction and method section on the Open Science Framework (see https://osf.io/vec5d/) before starting with data collection and data analysis to be able to distinguish between confirmatory and exploratory analyses, (e) hereby allow other researchers to re-analyze our data using our data files (including our entire literature hits from databases in common file formats) and statistical scripts, and (f) recruited expertise.

We followed the model of the APA by Tolin et al. (2015) by trying to classify the evidential value of our results as *very strong*, *strong*, or *weak*. For this classification, we applied the recommended GRADE system (Atkins et al., 2004; Guyatt et al., 2008) to assess the quality of our meta-analysis. A PICOTS approach was used to clearly define the population of interest (P), the intervention (I), the comparisons considered (C), the outcomes examined (O), the timing of outcome assessment (T), and the setting of treatment (S). However, a complete review process of LTPP according to Tolin et al. (2015) was not feasible in our study because our focus lay on RCTs and we did not review naturalistic, uncontrolled studies, which would be required for a full evaluation.

Definition of Long-Term Psychoanalytic Psychotherapy

We followed the definition of LTPP given in a previous meta-analysis by Leichsenring and Rabung (2011), but differed in the definition of the duration and dosage. Even though some experts regard long-term as lasting for 50 or more sessions (Crits-Christoph & Barber, 2000), we follow a current expert recommendation by Gabbard (2017) arguing that therapies lasting for 40 or more sessions should be included in the overarching rubric of long-term treatment. In accordance with Smit et al. (2012), we defined the duration and dosage of LTPP as lasting for at least 1 year *and* 40 sessions (in contrast to 1 year or 50 sessions in Leichsenring & Rabung, 2011). Additionally, we clearly distinguished between psychoanalysis and LTPP by considering the setting (de Maat et al., 2013). We, therefore, only included intervention conditions in which the therapist and the patient are in a sitting position.

Inclusion Criteria

In the first step of our study, we analyzed the same set of studies which Leichsenring et al. (2013) had included in their meta-analysis in order to reproduce their effect sizes and results. In our updated meta-analysis, we applied the following inclusion criteria according to the PICOTS approach:

(1) Patients/Problems: A clearly delineated sample of patients (≥ 18 years of age) with complex mental disorders (chronic mental disorders, more than one mental disorder, or personality disorder) is examined.
(2) Intervention: LTPP meets the definition given above and lasts for at least 1 year and 40 sessions.
(3) Comparator: Active control treatments differing substantially from the intervention treatment must be applied (a short-term treatment or a different type of treatment).
(4) Outcomes: Reliable and valid outcome measures are used (see Outcome Measures).
(5) Timing of outcome assessment: We analyzed pre- and post-treatment assessments (post-treatment defined as the point in time when the longer one of the compared treatments was finished); follow-up assessments were analyzed separately if provided.
(6) Setting/Study design: Studies are randomized or quasi-randomized (e.g., randomization by alternation or date of birth) controlled trials of individual therapy.
(7) Additionally, indicated data must allow to determine between-group effect sizes.

Information Sources and Search

We chose the online databases EBSCO (including MEDLINE, PsycINFO, PsycARTICLES), Web of Science, and Cochrane Library for our literature search. Document type was set to academic journals, journals, and dissertations for EBSCO, and articles, reviews, proceedings papers, and meeting abstracts for Web of Science. In Cochrane Library, search limits were set to all cochrane reviews, other reviews, and trials. We did not deviate in any other way from the default settings of these three databases. We used the same, well-tried search terms as Leichsenring and Rabung (2008, 2011): (psychodynamic OR dynamic OR psychoanalytic* OR transference-focused OR self psychology OR psychology of self) AND (therapy OR psychotherapy OR treatment) AND (study OR studies OR trial*) AND (outcome OR result* OR effect* OR change*) AND (psych* OR mental*) AND (rct* OR control* OR compar*). Additionally, we communicated with experts in the field to search for additional published as well as unpublished data. We had the opportunity to send a request for additional data via an international mailing list of around 500 psychoanalytic researchers. Additional data were accepted till June 11th, 2017. Besides, we manually screened reference lists in articles, reviews, and textbooks. Furthermore, all studies included in Leichsenring and Rabung (2011), Leichsenring et al. (2013), and Smit et al. (2012) were checked for inclusion in our meta-analysis.

Study Selection and Data Collection

The main author (CW) selected studies for inclusion. To remove ambiguity, the second author (FS) was consulted and consensus was reached through discussion. Data extraction was done in duplicate for all raw data. To assess the reliability of data extraction, we calculated Cohen's κ as recommended by Cooper (2010), keeping the limitations of this measure in mind (Eagan et al., 2017). Before our final analysis, consensus was reached through discussion for all codes.

We extracted data for the following variables: author names, publication year, date and place of the trial, publication status, psychiatric disorder of the sample, age and gender of patients, general clinical experience of therapists (years), specific experience with the patient group under study (years), duration of follow-up period, and co-interventions (e.g., use of psychotropic medication). The following variables were extracted for LTPP and the control treatment, respectively: type of treatment, duration, mean number of sessions, drop-outs, and mean number of sessions in completers. In a second data base, we collected all relevant data for effect size calculations (see Outcome Measures).

Assessment of the Risk of Bias in Individual Trials

To assess the risk of bias in individual trials, we used the Cochrane Risk of Bias Tool (Higgins et al., 2011). According to the tool, we assigned a judgement of *low*, *high,* or *unclear*

Table 1. Summary assessments of risk of bias within and across studies (adapted from Higgins et al., 2011)

Risk of bias	Interpretation	Within trial	Across trials
Low risk of bias	Bias, if present, is unlikely to alter the results seriously	Low risk of bias for all key items	The majority of trials carry a low risk of bias
Unclear risk of bias	A risk of bias raises some doubt about the results	Low or unclear risk of bias for all key items	The majority of trials carry a low or unclear risk of bias
High risk of bias	Bias may alter the results seriously	High risk of bias for one or more key items	The majority of trials carry a high risk of bias

risk of bias to our key items random sequence generation, allocation concealment, blinding of outcome assessment, incomplete outcome data, and selective reporting. The item blinding of participants and personnel was rated as unclear for all studies as considered by Higgins et al. (2011). Blinding of participants and personnel is impossible in psychotherapy research, but outcomes may be influenced by participants knowledge of which intervention they received. Hence, all studies were treated as being at a possible risk of bias on this item.

The following additional biases arose in our research process: (a) ongoing treatments were included, (b) data to calculate the exact effect size and/or session ratio were incomplete, which made an approximation necessary, and (c) therapies in the intervention and control group were carried out by the same therapists. We included these biases as key items.

If information concerning a key item was unavailable in a primary study, we rated the item as unclear. Mainly following an example by Higgins et al. (2011), we summarized the assessment of risk of bias within and across trials as shown in Table 1.

Since we a priori expected a small number of studies in our meta-analysis, we chose Higgins et al.'s (2011) recommended strategy: Present a meta-analysis of all studies and include the summary of the risk of bias across studies in the interpretation of the results.

For the sake of relevance, we used six out of eight quality criteria (two criteria are covered by the Cochrane Risk of Bias Tool) proposed by Cuijpers, van Straten, Bohlmeijer, Hollon, and Andersson (2010) to assess important aspects of psychotherapy research. We checked if (a) patients were diagnosed using a diagnostic system, (b) a treatment manual was used, (c) treatment integrity was checked (supervision or analysis adherence), (d) therapists were trained for intervention under study, (e) an intention-to-treat analysis was included, and (f) adequate statistical power and $n \geq 50$ were given. Additionally, we documented whether clinically significant change was measured and reported as recommended by Tolin et al. (2015). The exact coding rules for each item are provided in our uploaded excel file. Higgins et al. (2011) argue not to generate summary scores comprising *different* quality criteria across studies, which is why we solely assessed the *single* quality criteria across studies and included the findings in the interpretation of our results.

Outcome Measures

Following Leichsenring et al. (2013), we assessed effect sizes for the domains *general psychiatric symptoms*, *personality functioning*, *social functioning*, *target problems*, and *overall effectiveness*. Leichsenring et al. (2013) argue that recovery rates, which Smit et al. (2012) used, do not seem a reliable measure across studies of LTPP because definitions and measures of recovery are still too heterogeneous.

To assess general psychiatric symptoms, we searched the studies for broad symptom checklists such as the Symptom Checklist 90 (SCL-90; Derogatis, 1977) or other more specific symptom measures (e.g., measures of depression or anxiety as well as direct or indirect measures of single symptoms such as suicide attempts or hospitalization/emergency room visits for borderline patients). Personality functioning was assessed by measures of personality structure and personality characteristics focusing on the individual patient such as the Reflective Function Scale (Fonagy, Steele, Steele, & Target, 1998). For the assessment of social functioning, we used measures concerning a patient's social, interpersonal environment (e.g., the Social Adjustment Scale; Weissman & Bothwell, 1976, or measures assessing work ability or quality of life). To measure target problems, specific symptom measures (e.g., measures of depression for depressed patients, or measures of impulse control for borderline patients), direct and indirect measures of single symptoms, as well as measures of personality and social functioning which were specific to the patient group under study (e.g., measures of personality structure for borderline patients) were included. We first assigned measures to one of the three categories *general psychiatric symptoms*, *personality functioning,* or *social functioning*. Specific symptom measures, measures of single symptoms as well as measures of personality and social functioning which were specific to the patient group under study could then be additionally included in the domain *target problems* in order not to narrow the data basis in the respective outcome domain. Target problems are, thus, not independent of the three other domains.

In the overarching outcome measure *overall effectiveness*, we, therefore, only included the three independent domains general psychiatric symptoms, personality functioning, and social functioning by averaging the effect sizes of these domains. We excluded outcome measures for which the direction of the effect could not be clearly defined as unidirectional (e.g., fewer medication use can be generally regarded as an improvement, but for patients suffering from a bipolar disorder, for example, medication compliance may be regarded as a therapeutic target). The first author (CW) assigned each primary outcome measure to an outcome category. The first half of assignments was matched with a second rater of our working group to discuss and specify the definition of our outcome measures. We then calculated Cohen's κ for the second half of assignments by the two raters. Before our final analysis, consensus was reached through discussion for all assignments.

If a study contained more than one outcome measure for one domain, we assessed the effect size and variance separately for each measure and then calculated the mean effect size and mean variance for these measures. We averaged the variances because we did not know the correlation between the respective outcomes, which constitutes the conservative approach because one implies a correlation of 1.00 leading to an overestimation of the variance and an underestimation of precision (Borenstein, Hedges, Higgins, & Rothstein, 2009). However, this approach also assumes that the underlying true effects are homogeneous, which is not a conservative assumption. Hence, in case of heterogeneity, our effect sizes should be interpreted cautiously. If a study contained more than one intervention or control group, we preferred an outpatient individual LTPP intervention group to an inpatient LTPP intervention group because more trials deal with outpatients. Thus, we tried to reduce heterogeneity. If a study contained more than one outpatient intervention group, we assessed the effect size and variance separately for each group and then calculated the mean effect size and mean variance. Notably, this approach makes an implicit assumption of homogeneity which may lead to inaccuracies in case of heterogeneous interventions. We, therefore, tried to strictly follow our definition of LTPP with the goal to reduce heterogeneity. In the case of more than one control group, we chose (1) an evidence-based treatment (e.g., cognitive behavior therapy, interpersonal psychotherapy, or short-term psychoanalytical psychotherapy) over (2) a structured, non-evidence-based treatment with the most similar treatment intensity over (3) a structured, non-evidence-based treatment with the most similar treatment mode (outpatient or inpatient therapy, individual or group therapy) over (4) treatment as usual (TAU) or another non-structured treatment. We are aware of the fact that the approach of selecting the strongest comparator wastes information because data of the excluded comparators are not considered. However, by using this approach we tried to follow the APA recommendations by Tolin et al. (2015) to compare a psychotherapeutic intervention to the strongest comparator possible.

Since effect sizes may be overestimated if only patients who completed the treatment are included in the analysis, we collected *intention-to-treat (ITT)* data, in contrast to completers' data, where available (Hollis & Campbell, 1999). The concept of ITT analyses is that all randomized patients must be included in the analyses, no matter what happens after randomization (e.g., drop-out, detection of false inclusion). Exclusion must be very well justified. Common strategies for ITT analyses are carrying forward the last observed response or conservatively setting the effects for drop-out patients to zero. We followed Leichsenring and Rabung (2011) by choosing the latter strategy for studies which did not present ITT data. They adjusted a reported completers sample by multiplying a pre-post treatment difference of 0.5, for example, by the ratio of patients who completed the study and all included patients. With a completers' sample of 80 patients and 20 patients who dropped out of the study, the adjusted pre-post difference for the ITT analysis would be $0.4(0.5 \cdot \frac{80}{100})$.

Furthermore, we assessed effect sizes separately for post-treatment and the longest available follow-up because there is some empirical evidence that, after psychoanalytic psychotherapy was finished, psychotherapeutic gains may not only be maintained, but continue to improve (Town et al., 2012).

Statistical Analysis

We quantified between-group effect sizes by calculating standardized group mean differences bias-corrected for small sample sizes (i.e., Hedges' *g*; Hedges & Olkin, 1985). To calculate Hedges' *g* and its associated 95% confidence interval (CI), we collected means, sample standard deviations, and sample sizes for each outcome in each study. Our specific calculation was to subtract the mean pre-treatment to post-treatment (or follow-up) difference of the control condition from the respective difference of LTPP. This difference is divided by the pooled pre-treatment standard deviation and multiplied by a coefficient correcting for small sample size (Morris, 2008). We chose this approach in contrast to the standard Hedges' *g* calculation (see Table 2 for the different calculations) in order to show the difference in pre-post changes of the compared groups which, thus, comprises more information than the standard calculation. Pre-post correlations are needed for the exact computation of the variance for this outcome measure, but were unavailable for virtually all studies. Only one study

Table 2. Overview of the three different outcome metrics: Standard Hedges' *g*, pre-post-control Hedges' *g*, and complemented pre-post-control Hedges' *g*

Metric	Research question	Abbreviation	Formula
Standard Hedges' *g*[a]	RQ 1: Do participants of the long-term group have better outcomes at the end of the therapy than participants of the control group (ignoring potential group differences at the beginning of the therapy)?	Standard *g*	(1) $c_{\mathrm{a}} \times \frac{M_{\mathrm{post,C}} - M_{\mathrm{post,T}}}{SD_{\mathrm{post}}}$
Pre-post-control Hedges' *g*	RQ 2: Do participants of the long-term group have a stronger increase of positive outcomes from the beginning of the therapy to the end, compared to the increase of participants in the control group (i.e., taking potential group differences at the beginning of the therapy into account)?	ppc-*g*	(2) $c_b \times \frac{(M_{\mathrm{pre,T}} - M_{\mathrm{post,T}}) - (M_{\mathrm{pre,C}} - M_{\mathrm{post,C}})}{SD_{\mathrm{pre}}}$
Complemented pre-post-control Hedges' *g*[b]	Same as RQ 2, but complementing 46 outcomes without pre-measurements (out of 191 outcomes), which are missing in RQ 2, by their post group comparison as in RQ 1 (primarily provided for a comparison with Leichsenring et al., 2013)	compl. ppc *g*	Formula (2); formula (1) if pretest outcomes were not presented

Notes. $c_{\{a,b\}}$ = coefficient correcting for small sample size; *M* = means of the pre-treatment (pre, T) and pre-control (pre, C) as well as post-treatment (post, T) and post-control (post, C) outcomes; SD_{post} = the pooled post-treatment standard deviation; SD_{pre} = the pooled pre-treatment standard deviation. [a]Metric applied by Smit et al. (2012). [b]Our main outcome measure, also applied by Leichsenring et al. (2013).

(Levy et al., 2006) reported pre-post correlations for four outcome measures (*r*s = .22, .56, .72, .82). Given this range of pre-post correlations, we generally plugged in a default correlation of .5. If means and standard deviations are not reported in a primary study, but the *t*-test or *F*-test associated with the difference in post means is presented, it is also possible to calculate Hedges' *g* from these metrics (Cooper, 2010). In the following, we will refer to our approach of calculating Hedges' *g* as pre-post-control (ppc) *g* complemented by calculations from other metrics as well as standard Hedges' *g* calculations if pre-means were not presented (i.e., compl. ppc-*g*). We complemented the ppc-*g* in order not to narrow the data basis, since we expected a small set of studies. Notably, with an increasing data base future meta-analysts should not combine these metrics since they test slightly different hypotheses. To provide a robustness check for our approach, we still reported results from standard Hedges' *g* calculations and ppc-*g* with no complementation. We decided to provide these different calculations because the previous meta-analyses by Smit et al. (2012) and Leichsenring et al. (2013) differed in their calculations and we wanted to compare our results to both meta-analyses. Smit et al. (2012) used the standard Hedges' *g*, whereas Leichsenring et al. (2013) used the compl. ppc-*g*.

In our case, a positive effect size indicated improvement. Signs were inverted if necessary. We asked the original authors for necessary data if not reported. If only the overall sample size was reported, we assumed sample sizes to be equal across groups (in case of an odd total sample size, we placed the remainder in the LTPP group).

The effect sizes were aggregated using a random-effects model because we expected some dispersion or heterogeneity in observed effects due to different types of disorders and a variety of control treatments (Borenstein et al., 2009). Meta-analytical heterogeneity was assessed by using the χ^2, *Q* statistic, the I^2 statistic, and the parameter τ^2 (Borenstein et al., 2009). To directly communicate the amount of heterogeneity in the underlying true effects, we present the 95% prediction intervals (PI) for each outcome measure (IntHout, Ioannidis, Rovers, & Goeman, 2016). Prediction intervals show the expected range of true effects in similar studies. These intervals are broader than the 95% CIs as they additionally take heterogeneity into account. We explored heterogeneity by considering the session ratio as a possible moderating variable. The relation between session ratio and effect size was examined by meta-regression. We log-transformed the ratios, so that they are symmetric around 1.

To assess and adjust for publication bias, we followed Carter, Kofler, Forster, and McCullough (2015) by conducting the Egger test (Egger, Smith, Schneider, & Minder, 1997), the Precision Effect Test (PET), and the Precision Effect Estimation with Standard Error (PEESE; Stanley and Doucouliagos, 2014). Additionally, we applied the *p*-uniform approach (van Assen, van Aert, & Wicherts, 2015), as well as a three-parameter selection model (3PSM; McShane et al., 2016). First, we examined funnel plot asymmetry by conducting the Egger test as well as *p*-uniform's test for publication bias, and by determining the selection parameter of 3PSM. In the presence of publication bias, a funnel plot (i.e., a plot of trials' effect sizes against their precision) is skewed and asymmetrical with more studies on the right side of an inverted funnel. To quantify the relation of reported effect sizes to its standard error, Egger et al. (1997) used a linear approximation. By applying a weighted least squares (WLS) regression model in which trials' effect sizes are regressed on the respective standard error, weighted by the inverse of the variances, publication bias may be quantified as the slope coefficient – for this model (Stanley & Doucouliagos, 2014). Testing H_0: α = 0 provided us with the information whether publication bias was present or not. The significance level was set at .10 as recommended by Egger et al. (1997). Additionally, testing H_0: γ = 0, with γ as the intercept in this WLS regression model, has been proposed as a test for a real empirical effect beyond publication bias. Stanley and Doucouliagos (2014) proposed that the coefficient γ of this model may serve as an estimate of the effect size adjusted for publication bias, which is called PET. PET provides an attempt to correct for publication bias, but according to recent simulation studies not a very accurate one (Carter, Schönbrodt, Hilgard, & Gervais, 2017). PEESE works with the same method, but uses variances instead of standard errors as the predictor in the WLS regression model. Stanley and Doucouliagos (2014) argued to apply PET and PEESE regardless of the statistical significance of the Egger test because the Egger test notoriously has low power. Furthermore, their simulation studies showed that in the case of a genuine effect, PET tends to underestimate if an underlying effect exists, meaning that it overcorrects for the influence of publication bias. In case of no underlying effect, PEESE tends to overestimate the effect, meaning that it undercorrects for publication bias. Therefore, Stanley and Doucouliagos (2014) recommended using the estimate of the effect size given by PEESE if PET is statistically significant, and using the estimate given by PET if PET is not statistically significant. We report both estimates to provide a thorough documentation.

Additionally, we applied the *p*-uniform approach by van Assen et al. (2015) to test and correct for publication bias. The distribution of statistically significant *p*-values ($p < .05$) across a set of studies is called *p*-curve. The shape of the distribution is uniform if there is no effect, and right-skewed if there is one. For a given set of statistically significant *p* values, it is, thus, possible to diagnose the presence of publication bias and to compute an effect size estimate that corrects for publication bias. The *p*-curve approach by Simonsohn, Nelson, and Simmons (2014) is based on the same basic idea, but simply differs in implementation. Since both approaches basically yield the same results, we decided to only apply *p*-uniform because this method also provides confidence intervals.

Our final approach to assess and adjust for publication bias were two selection methods (Hedges, 1984; McShane et al., 2016). Selection methods have two components: A data model and a selection model. The data model characterizes how the data are generated, whereas the selection model characterizes the publication process. Selection models can adapt to different biases in the publication process, such as (a) only studies with statistically significant results are published, (b) only studies with statistically significant and directionally consistent results are published, or (c) studies with non-significant results (or results which are directionally inconsistent) are less likely to be published than studies with significant and directionally consistent results. One-parameter approaches, such as described in Hedges (1984), *p*-curve, and *p*-uniform, assume that (a) only studies with statistically significant results are published and (b) effect sizes do not vary across studies (i.e., they are homogeneous). Since both assumptions are almost always false in behavioral research, McShane et al. (2016) showed in a large simulation study that a three-parameter selection model (3PSM) provides better estimates and tests. Such a selection estimates (a) the underlying effect size adjusted for publication bias, (b) the degree of heterogeneity adjusted for publication bias, and (c) the degree of publication bias, which is formalized as a weight parameter, that provides the probability that a study with statistically non-significant results is published relative to a study with statistically significant results. Notably, our averaging of multiple outcomes may violate the assumptions of *p*-uniform and selection modeling since these methods assume that outcomes are published on the basis of their individual *p*-values, rather than the composite *p*-value of averaged outcomes.

In case methods disagreed about the presence and/or magnitude of a bias-corrected meta-analytical effect, we pre-registered to give 3PSM the largest weight, as this method, among other reasons, had consistently a better performance in many conditions than other bias-correcting methods (Carter et al., 2017; McShane et al., 2016).

Furthermore, we screened for outliers in an exploratory manner and conducted sensitivity analyses where necessary. Sensitivity analyses and all bias-correcting methods were only conducted for our primary outcome, meaning the post-treatment compl. ppc-*g* uncorrected for ITT. The data basis for the follow-up assessment was too narrow and the post-hoc correction for ITT, where ITT data were not reported in the primary studies, can only be considered as a very rough estimation. Fortunately, most of the primary studies reported ITT data which made a post-hoc correction unnecessary for these studies.

Post hoc, we decided to include a four-parameter selection model (4PSM) for further exploratory analyses. The fourth parameter, in other words, the second selection parameter, provides the likelihood that a study with a negative effect size is published. We used *p*-value bin thresholds at a one-sided *p*-value of .025 and .50.

Our statistical analyses were conducted using R (version 3.4.0; R Development Core Team, 2008). Data were aggregated using the *metafor* package (Viechtbauer, 2010). For the assessment of publication bias, we additionally used the *puniform* package (van Aert, 2017; van Aert, Wicherts,

Table 3. Comparing Leichsenring et al.'s (2013) parameter estimates of the comparison of long-term psychoanalytic psychotherapy (LTPP) versus shorter forms of psychotherapy to our replicated estimates

Domain	Comparison	*k*	*g*	95% CI	*Q*
Psychiatric symptoms	Leichsenring et al. (2013)	11	0.43	[0.20, 0.67]	13.60
	Replication	12	0.28	[0.09, 0.47]	22.12*
Target problems	Leichsenring et al. (2013)	12	0.39	[0.06, 0.72]	13.36
	Replication	12	0.31	[0.08, 0.54]	28.66**
Social functioning	Leichsenring et al. (2013)	9	0.60	[0.23, 0.97]	14.01†
	Replication	10	0.43	[0.14, 0.72]	31.16***
Personality functioning	Leichsenring et al. (2013)	8	0.41	[−0.18, 1.00]	6.25
	Replication	7	0.45	[0.18, 0.71]	9.69
Overall effectiveness	Leichsenring et al. (2013)	12	0.40	[0.05, 0.74]	14.01
	Replication	12	0.35	[0.16, 0.53]	21.75*

Notes. All data were calculated using a random-effects model. *k* = number of comparisons; *g* = estimate of the average underlying effect (pre-post-control Hedges' *g* complemented by calculations from *t*-values and other metrics as well as by standard Hedges' *g* calculations); CI = the upper and lower limits of the 95% confidence interval; *Q* = *Q* statistic for statistical heterogeneity. $^{\dagger}p < .1$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

& van Assen, 2016) and the *weightr* package (Coburn & Vevea, 2017).

Changes to the Pre-Registration

We did not differ from our pre-registration in any major way. We present all minor changes in Electronic Supplementary Material (ESM 1).

Results

Replication of Leichsenring et al. (2013)

For the replication of Leichsenring et al. (2013), we included the same set of 12 studies they had included (See Table 4, but excluding the more recent studies (7), (11), and (13)). For the replication we also included the study by Dare, Eisler, Russell, Treasure, and Dodge (2001) which did not fulfill our duration and dosage criteria for our updated meta-analysis. Table 3 summarizes the comparison between the outcomes of Leichsenring et al. (2013) and our replication for the five outcome domains.

Our classification of outcome domains differed for the study by Giesen-Bloo et al. (2006). We could not clarify with Falk Leichsenring whether he and his team had included data from this article or from a second article of this study (van Asselt et al., 2008). We decided to include data from the latter article because it reported means instead of medians and more treatments were finished than at the time of the first article. In accordance with the definition of the outcome domains, both authors and an independent rater clearly agreed to assign the outcome measures of this study to the domain psychiatric symptoms (the Borderline Personality Disorder Severity Index, BPDSI) and to the domain social functioning (the utility scores), whereas Leichsenring et al. (2013) assigned both outcome measures to the domain personality functioning. The utility scores capture dimensions such as mobility, self care, and daily activities, which is why we clearly assigned them to social functioning. The assignment of the BPDSI can be controversially discussed. The BPDSI is an indicator of borderline symptoms and represents the DSM-IV borderline personality disorder criteria, which can be considered to reflect personality functioning because they describe a personality disorder. However, we wanted to distinguish between measures which focus more on evaluating the diagnostic criteria, symptoms or severity of a disorder, which we assigned to the category of psychiatric symptoms, and measures which focus more on personality structure and structural representations, such as the Reflective Function Scale (Fonagy et al., 1998), which we assigned to the category of personality functioning. It was difficult to draw the line for measures of borderline personality disorder, because the criteria of personality disorder reflect personality functioning. Compared to the Reflective Function Scale, we (and an independent second coder) still considered the BPDSI rather to be a measure of psychiatric symptoms. The discrepancy between Leichsenring et al.'s (2013) and our assignment explains the differing number of comparisons for the outcome domains psychiatric symptoms, social functioning and personality functioning.

Consistent with the meta-analysis by Leichsenring et al. (2013), we found LTPP to be significantly superior to comparators in all outcome domains. However, our effect sizes were smaller by 0.15, 0.08, 0.17, and 0.05 for the outcome domains psychiatric symptoms, target problems, social functioning, and overall effectiveness, respectively. For personality functioning, we received a slightly higher effect size (by 0.04). Our 95% CIs were narrower for all outcome domains. The *Q* statistic provided only one statistically significant result for the outcome domain social functioning

Table 4. Main characteristics of included studies of long-term psychoanalytic psychotherapy

ID	Study	Disorder	LTPP intervention	Control intervention	n_{LTPP}	n_{CONTROL}	SR
(1)	Bachar et al. (1999)	Bulimia and anorexia	Self psychological treatment	Cognitive orientation treatment	17 (3)	17 (5)	1.0
(2)	Bateman and Fonagy (1999)	BPD	Mentalization-based therapy	General psychiatric outpatient care	22 (3)	22 (3)	1.8
(3)	Bateman and Fonagy (2009)	BPD	Mentalization-based therapy	Structured clinical management	71 (19)	63 (16)	NA
(4)	Bressi et al. (2010)	Anxiety or depressive disorder	Psychodynamic psychotherapy	Drug treatment and clinical interviews	30 (6)	30 (6)	1.6
(5)	Clarkin et al. (2007)	BPD	Transference-focused psychotherapy; dynamic supportive treatment	Dialectical behavior therapy	61 (16)	29 (12)	1.0
(6)	Doering et al. (2010)	BPD	Transference-focused psychotherapy	Experienced community psychotherapy	52 (20)	52 (35)	2.6
(7)	Fonagy et al. (2015)	Depressive disorder	Psychodynamic psychotherapy	Treatment as usual	67 (16)	62 (16)	3.7
(8)	Giesen-Bloo et al. (2006)	BPD	Transference-focused psychotherapy	Schema-focused therapy	42 (21)	44 (11)	1.2
(9)	Gregory et al. (2008)	BPD and alcohol use disorder	Dynamic deconstructive psychotherapy	Treatment as usual	15 (5)	15 (6)	0.6
(10)	Huber, Zimmermann, et al. (2012)	Depressive disorder	Psychodynamic psychotherapy	Cognitive behavior therapy	35 (5)	41 (10)	2.0
(11)	Jørgensen et al. (2013)	BPD	Mentalization-based therapy	Supportive group treatment	58 (16)	27 (6)	4.0
(12)	Knekt, Lindfors, Härkänen, et al. (2008)	Anxiety and depressive disorder	Psychodynamic psychotherapy	Short-term psychodynamic psychotherapy	128 (47)	101 (13)	5.0
(13)	Poulsen et al. (2014)	Bulimia	Psychoanalytic psychotherapy	Cognitive behavior therapy	34 (10)	36 (8)	3.3
(14)	Svartberg et al. (2004)	Cluster C personality disorder	Psychodynamic psychotherapy	Cognitive therapy	26 (1)	25 (0)	1.0

Notes. Only the main article of a study is listed here. LTPP = long-term psychoanalytic psychotherapy; *n* = sample size (drop-outs); SR = session ratio of the mean number of sessions of LTPP versus the mean number of sessions of the control group; BPD = borderline personality disorder; NA = not available.

in Leichsenring et al.'s (2013) meta-analysis. In our replication, statistical heterogeneity was statistically significant for all outcome domains except personality functioning.

Falk Leichsenring supplied us with their aggregated effect size data of the single studies for the single domains. The 51 single comparisons across all studies and all outcome domains are shown in Table 1 in ESM 1. Of course this comparative table cannot tell which of the extracted effect sizes are the more appropriate, but it highlights the difficulties of producing reproducible meta-analyses (Lakens et al., 2017).

To sum up, we replicated the direction and general tendency of results, but our replicated effect sizes were in general slightly smaller and showed higher heterogeneity. The effect size for personality functioning was slightly higher.

Updated Meta-Analysis

Study and Outcome Selection

For our updated meta-analysis, we screened a total of 9,170 records. We excluded 9,144 of them. Figure 2 illustrates our search, screening, and selection process. Studies were mainly excluded because the intervention did not meet our definition of LTPP or the trial was not randomized and controlled. Worth mentioning are the studies by Linehan et al. (2006) and McMain et al. (2009) because they were included in the meta-analysis by Smit et al. (2012). We agree with Leichsenring et al. (2013) that no LTPP group was examined in these studies. Three studies were excluded because they were not randomized and/or not clearly controlled (Klar, 2005; Korner, Gerull, Meares, & Stevenson, 2006; Puschner, Kraft, Kächele, & Kordy, 2007). We did not include two studies because they did not meet the dosage or the duration criteria for our definition of LTPP (Dare, Eisler, Russell, Treasure, & Dodge, 2001; Zipfel et al., 2014).

We found 14 studies described in 26 articles meeting our inclusion criteria. We did not receive any unpublished or additional data from researchers in the field. The 14 main articles are: Bachar, Latzer, Kreitler, and Berry (1999), Bateman and Fonagy (1999, 2009), Bressi, Porcellana, Marinaccio, Nocito, and Magri (2010), Clarkin, Levy, Lenzenweger, and Kernberg (2007), Doering et al.

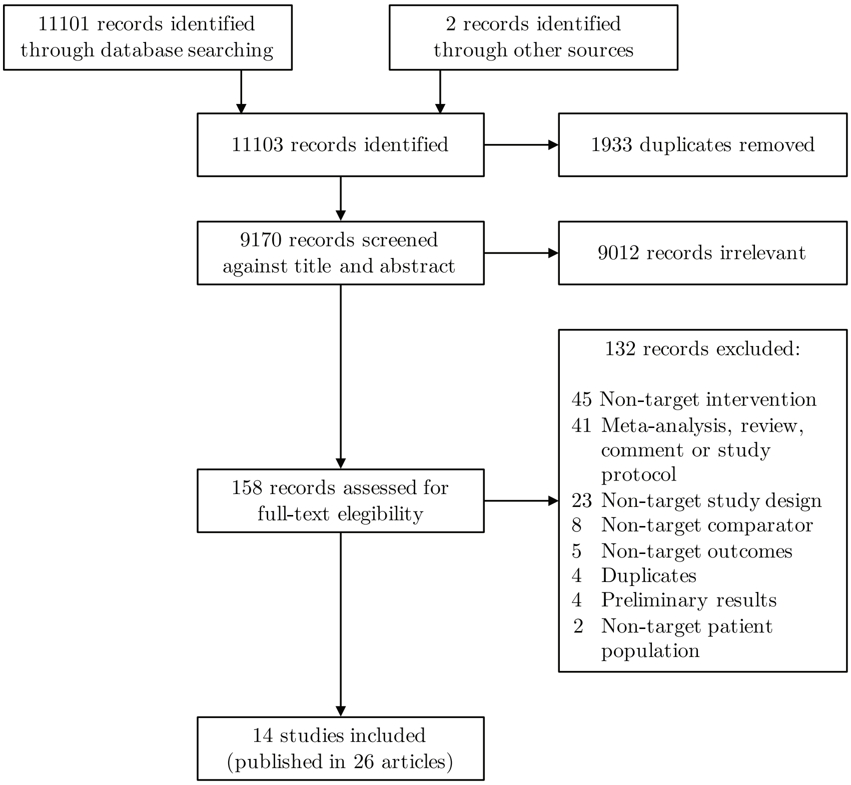
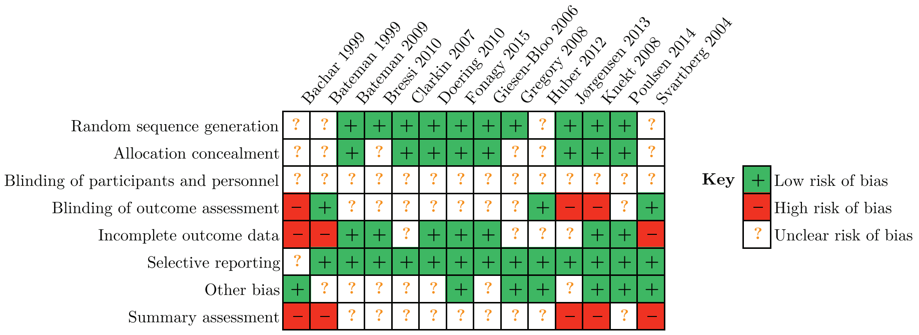

11101 records identified through database searching
2 records identified through other sources
11103 records identified
1933 duplicates removed
9170 records screened against title and abstract
9012 records irrelevant
132 records excluded:
45 Non-target intervention
41 Meta-analysis, review, comment or study protocol
23 Non-target study design
8 Non-target comparator
5 Non-target outcomes
4 Duplicates
4 Preliminary results
2 Non-target patient population
158 records assessed for full-text elegibility
14 studies included (published in 26 articles)

Figure 2. Flowchart of study search, screening and selection. Figure available at https://osf.io/vec5d/, under a CC-BY 4.0 license.

(2010), Fonagy et al. (2015), Giesen-Bloo et al. (2006), Gregory et al. (2008), Huber, Zimmermann, et al. (2012), Jørgensen et al. (2013), Knekt, Lindfors, Härkänen, et al. (2008), Poulsen et al. (2014), and Svartberg, Stiles, and Seltzer (2004). Additional outcome measures or follow-up data were reported in the following 12 articles: Bateman and Fonagy (2008), Gregory, DeLucia-Deranja, and Mogle (2010), Huber, Henrich, Clarkin, and Klug (2013), Huber, Henrich, Gastner, and Klug (2012), Jørgensen et al. (2014), Knekt et al. (2015), Knekt, Lindfors, Laaksonen, et al. (2008), Knekt et al. (2016), Levy et al. (2006), Lindfors, Knekt, Heinonen, Härkänen, and Virtala (2015), Lindfors, Knekt, Virtala, and Laaksonen (2012), and van Asselt et al. (2008).

Doering et al. (2010) and Giesen-Bloo et al. (2006) reported on ongoing treatments which is why we ran a sensitivity analysis without them. The study by Clarkin et al. (2007) did not report means and standard deviations. We contacted the authors twice, but did not receive a reply. Therefore, we calculated effect sizes based on the reported effect sizes *r* of their regression analyses. Being an approximation, we regarded this calculation as another risk of bias in our assessment of the risk of bias within studies. Furthermore, we discussed the inclusion of Jørgensen et al. (2013) because the intervention group and the control group were partly treated by the same therapists, which raises the question whether the difference between the two treatment arms was too narrow. Since the two treatment arms were well-described as applying combined mentalization-based therapy (one individual and one group session per week) in the treatment group and supportive group treatment (one group session biweekly) in the comparison group, we regarded the difference as sufficient enough and decided to include the study. It is noteworthy that the treatment group was confounded by a group session which is part of the mentalization-based treatment program. The same applied to the studies by Bateman and Fonagy (1999, 2009). With a hopefully increasing data basis, future meta-analysts should examine the effects of mentalization-based therapy in a subgroup analysis.

Some studies reported more than one intervention or control group. For the study by Clarkin et al. (2007), we combined the transference-focused psychotherapy group with the dynamic supportive treatment group to one intervention group because both treatments fulfilled our criteria of LTPP. Huber, Zimmermann, et al. (2012) examined three groups: Patients receiving (1) CBT, (2) LTPP, or (3) psychoanalysis proper. Leichsenring et al. (2013) combined the LTPP and the psychoanalysis proper groups, whereas we considered it to be more accurate according to our predefined inclusion criteria to exclude the psychoanalysis proper group and to solely examine the comparison of CBT versus LTPP. We chose the cognitive orientation treatment plus nutritional counselling group over the solely nutritional counselling group as control conditions of the study by Bachar et al. (1999), and the STPP group over



	Bachar 1999	Bateman 1999	Bateman 2009	Bressi 2010	Clarkin 2007	Doering 2010	Fonagy 2015	Giesen-Bloo 2006	Gregory 2008	Huber 2012	Jørgensen 2013	Knekt 2008	Poulsen 2014	Svartberg 2004
Random sequence generation	?	?	+	+	+	+	+	+	+	?	+	+	+	?
Allocation concealment	?	?	+	?	+	+	+	+	?	?	+	+	+	?
Blinding of participants and personnel	?	?	?	?	?	?	?	?	?	?	?	?	?	?
Blinding of outcome assessment	−	+	?	?	?	?	?	?	?	+	−	−	?	+
Incomplete outcome data	−	−	+	+	?	+	+	+	?	?	?	+	+	−
Selective reporting	?	+	+	+	+	+	+	+	+	+	+	+	+	+
Other bias	+	?	?	?	?	?	+	?	+	+	?	+	+	+
Summary assessment	−	−	?	?	?	?	?	?	?	?	−	−	?	−

Key	
+	Low risk of bias
−	High risk of bias
?	Unclear risk of bias

Figure 3. Summary assessment of the risk of bias within studies by applying the Cochrane Risk of Bias Tool (Higgins et al., 2011). Figure available at https://osf.io/vec5d/, under a CC-BY 4.0 license.

the solution-focused therapy group of Knekt, Lindfors, Härkänen, et al. (2008).

Study Characteristics

Table 4 presents the main characteristics of the 14 included studies. In total, the 14 included studies encompassed 658 patients who received LTPP and 564 patients who were treated with comparative treatments. The mean number of sessions across all studies was 81.8 (*SD* = 66.1) for the LTPP condition and 49.0 (*SD* = 45.6) for the treatments in the control groups, implying that we compared LTPP primarily to other forms of long-term psychotherapy. The medians were 58 and 37.5, respectively. The overall session ratio was 1.67, implying that patients treated with LTPP received 1.67 times as many sessions as patients treated with the control treatments. Seven studies reported follow-up data, some in different articles (Bateman & Fonagy, 2008; Fonagy et al., 2015; Gregory et al., 2010; Huber, Zimmermann, et al., 2012; Jørgensen et al., 2014; Knekt et al., 2016; Svartberg et al., 2004). Since many treatments were ongoing in the study by Giesen-Bloo et al. (2006), we extracted data from the paper by van Asselt et al. (2008) and considered them as post-treatment data. More treatments were finished in this second paper but not all of them.

Assessment of the Risk of Bias in Individual Trials

Figure 3 summarizes our findings for each study across the items of the Cochrane Risk of Bias Tool. We detected a high risk of bias for five studies because these studies presented incomplete outcome data and/or did not blind the outcome assessors. The rest of the studies carried an unclear risk of bias, mainly because they did not report whether the intended blinding of outcome assessment was effective or not, a criterion which is required by the tool. According to our pre-defined summary assessment across trials (see Table 1), we concluded that our set of studies carried an unclear risk of bias, suggesting that the results of the following meta-analysis should be interpreted cautiously.

The assessment of the six quality criteria by Cuijpers et al. (2010) revealed the following results across the 14 studies for the single items: In 13 studies, diagnostic systems were used to diagnose patients; in 10 studies, treatment manuals were used, which may be regarded as a very positive result given the difficulties with manuals for long-term and especially psychoanalytic treatments; in 13 studies, treatment integrity was checked; in 13 studies, therapists were trained for the intervention under study; in 11 studies, ITT analysis was included; and in 6 studies, an adequate statistical power and a sample size larger than 50 were given. Besides, we identified only five studies which measured and reported clinically significant change as required by Tolin et al. (2015). Considering these quality criteria, we drew the conclusion that a high proportion of studies fulfilled five out of seven criteria, implying that our primary studies abided by the bigger part of required standards. The assessment of the criteria regarding statistical power and clinically significant change, however, revealed poor results.

Reliability of Our Coding

Cohen's κ was 0.97 for the assignment to the mutually exclusive outcome domains, and 1.00 for the additional assignment to the domain target problems (52% of the outcome measures were additionally included in the domain target problems). The (almost) perfect agreement between the two raters suggests a very thorough definition of our outcome measures. The reliability of our data extraction was substantial. Cohen's κ was 0.93 for the general data, 0.70 for the outcome data, and 0.73 for the additional metrics. All inconsistencies were resolved by discussion prior to data analysis.

Different Calculations of Hedges' *g* and Follow-Up Data

The summary effects of the post data revealed very similar values for the standard *g* and our primary outcome, the compl. ppc-*g*, suggesting that our robustness check of the compl. ppc-*g* as our primary outcome measure was positive (see ESM 1 for all values). The pure ppc-*g* calculations yielded to some extent higher results for all outcome

Table 5. Comparing long-term psychoanalytic psychotherapy (LTPP) with other forms of psychotherapy: Parameter estimates for the random-effects models of the main outcome

Domain	*k*	*g*	*SE*	95% CI	*Q*	τ^2	I^2	95% PI
Psychiatric Symptoms	14	0.24**	0.09	[0.06, 0.42]	26.71*	0.05	50.94	[−0.30, 0.78]
Target Problems	14	0.25*	0.12	[0.02, 0.48]	38.94***	0.13	70.99	[−0.57, 1.06]
Social Functioning	13	0.35**	0.12	[0.11, 0.59]	40.37***	0.14	73.24	[−0.50, 1.20]
Personality Functioning	10	$0.24^{\dagger}$	0.14	[−0.03, 0.51]	22.73**	0.11	65.47	[−0.58, 1.07]
Overall Effectiveness	14	0.28**	0.10	[0.09, 0.47]	29.58**	0.07	57.06	[−0.33, 0.89]

Notes. *k* = number of comparisons; *g* = estimate of the average underlying effect (pre-post-control Hedges' *g* complemented by calculations from t-values and other metrics as well as by standard Hedges' *g* calculations); *SE* = standard error of the estimate of the average underlying effect; CI = the upper and lower limits of the 95% confidence interval; *Q* = *Q* statistic for statistical heterogeneity; τ^2 = estimate of the between-study variance; I^2 = percentage of the observed variance which is due to real differences in effect sizes; PI = the upper and lower limits of the 95% prediction interval. $^{\dagger}p < .1$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

domains, implying, however, that our calculation constitutes a more conservative effect size estimation. The differences between ITT corrected values and the respective uncorrected values were negligible, except for the moderation model of personality functioning.

The follow-up data yielded to some extent higher effect sizes than the post data for most of the three different calculations (see ESM 1). These data should be interpreted cautiously because they are only based on seven studies. In the following, we will, therefore, only report on further results of our primary outcome, meaning compl. ppc-*g* uncorrected for ITT and based on the post-treatment data. We chose the outcomes uncorrected for ITT because the difference to the post-hoc corrected values was negligible and because a post-hoc correction can only be considered as a very rough estimation (see Method section).

Main Outcomes

Table 5 presents the parameter estimates for the random-effects models of the main outcome (posttest data of compl. ppc-*g*). For the outcome domains psychiatric symptoms, target problems and overall effectiveness, all 14 included studies offered data. The number of comparisons was by one smaller for social functioning because Bachar et al. (1999) did not include outcome measures concerning social functioning. Four studies did not include outcome measures of personality functioning (Bateman & Fonagy, 2009; Bressi et al., 2010; Fonagy et al., 2015; van Asselt et al., 2008).

The meta-analytic effect sizes were statistically significant for all outcome domains (*p*s < .05), except the domain personality functioning (*p* = .08). According to Cohen's (1988) benchmarks (0.2 < *d* < 0.5 for small effects, 0.5 < *d* < 0.8 for medium effects, *d* ≥ 0.8 for large effects), the sizes of the effects are regarded as small. The lower limits of the 95% CIs lay in the range of a negligible effect for all outcome domains, whereas the upper limits lay in the range of a medium effect size for the domains social functioning as well as personality functioning and were higher than 0.40 for the other domains. The *Q* statistic suggests effect size heterogeneity in all outcome domains.

Concerning the I^2 descriptor (Higgins & Green, 2008), the outcome domain psychiatric symptoms and overall effectiveness fell within the range of moderate heterogeneity (i.e., 30% < I^2 < 60%) and all outcomes fell within the range of substantial heterogeneity (i.e., 50% < I^2 < 90%). The estimates of psychiatric symptoms and overall effectiveness lay in the overlap of the two ranges suggesting a moderate to substantial heterogeneity. The moderate to substantial heterogeneity was also expressed by the wide 95% PIs, which included medium to large effects in favor of the LTPP group, but also small to medium effects in the reverse direction (i.e., in favor of the control group) for all outcome measures. The effect sizes of the single studies with their 95% CI are presented in Figure 4 for the outcome domain overall effectiveness. The funnel and forest plots for the other four domains are presented in the ESM 1. The studies varied in size and direction of the effect for all outcome domains with the majority of studies showing positive effect sizes.

To conclude, the random-effects models yielded significant but small positive effect sizes for all outcome domains, except the domain personality functioning (*p* = .08). Considering the moderate to substantial heterogeneity and the range of the 95% CI and 95% PI, these results should be interpreted cautiously.

Sensitivity Analysis

We conducted a sensitivity analysis excluding two studies of ongoing treatments (Doering et al., 2010; Giesen-Bloo et al., 2006). The differences of the effect sizes to our main outcome were very small (see ESM 1 for all parameters). We concluded that despite the inclusion of ongoing treatments, our results of the main outcome are robust. However, in accordance with Leichsenring et al.'s (2013) argumentation and findings that the inclusion of ongoing treatments yields smaller effect sizes, our effect sizes of the main outcome were also slightly smaller for the

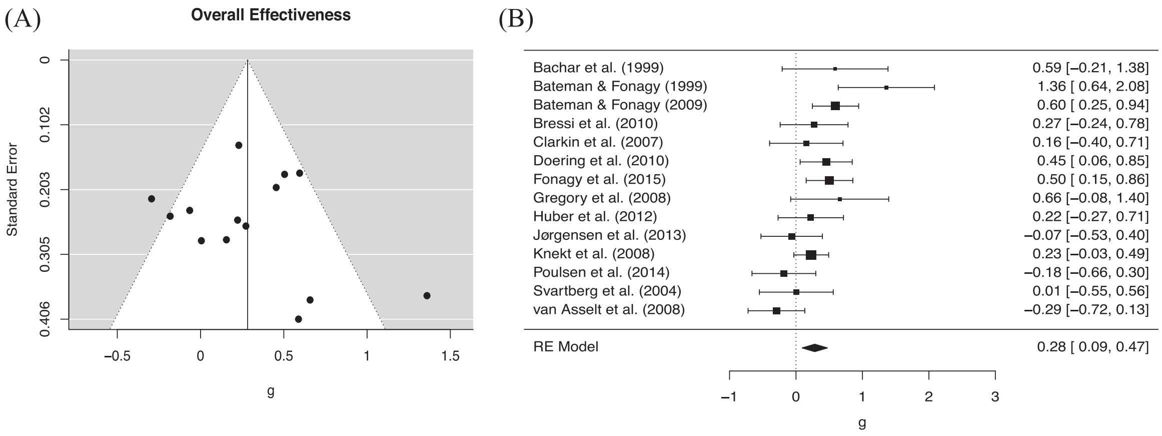


Figure 4. Funnel plot (A) and forest plot (B) for the outcome domain overall effectiveness. In the funnel plot, the standard error is presented on the inverted vertical axis and the standardized mean difference (*g*, in our case meaning the pre-post-control g complemented by *t*-values and other metrics as well as standard Hedges' *g* calculations) on the horizontal axis. Each dot represents one study. The inverted funnel is centered on the random-effects meta-analysis estimate of LTPP. The forest plot shows the effect size g with its associated 95% confidence interval for each study and for the summary effect at the bottom. The boxes show the effect size and the size of the boxes represents the relative weight. Figure available at https://osf.io/vec5d/, under a CC-BY 4.0 license.

domains psychiatric symptoms, target problems, social functioning and overall effectiveness compared to those of the sensitivity analysis.

Besides, we visually screened the funnel plots of all outcome domains for outliers which we defined as studies clearly lying outside of the inverted funnel, meaning not touching the inverted funnel. Removing the outliers (see ESM 1for an exact description of each outcome domain), our results only differed very slightly in the size of the effect for the domains psychiatric symptoms, target problems, social functioning and overall effectiveness. However, our marginal significant result for the domain personality functioning was not robust to our sensitivity analysis.

Moderation by Session Ratio

The outcomes of the meta-regression for our main outcome (compl. ppc-*g*) revealed that the session ratio was only associated with personality functioning. The regression coefficients for the intercept and the slope of the mixed-effects model were statistically significant (b_0 = 0.52, p = .004; $b_1 = -0.39$, p = .03). However, these results were not in line with our hypothesis that an increasing session ratio leads to an increasing effect size. In contrast, with a negative slope the results point in the opposite direction, meaning that the larger the dosage of LTPP compared to the control group is, the smaller the effect size is. These results should be interpreted very cautiously because the model was solely based on 10 studies. Furthermore, the *p*-value of the slope would not have survived a correction for multiple testing. No association between the session ratio and the effect size was found for the other outcome domains (*p*-values for the intercepts and slopes were > .07 and > .63, respectively).

Assessment of and Adjustment for Publication Bias

To assess publication bias, we first examined funnel plot asymmetry by conducting the Egger test as well as *p*-uniform's test for publication bias and by determining the selection parameter of 3PSM and 4PSM. To illustrate the association of effect size and precision, the funnel plot for the domain overall effectiveness is presented in Figure 4. The Egger test only revealed a statistically significant result for the domain personality functioning (p = .03). Publication bias for the domains psychiatric symptoms, target problems, social functioning, and overall effectiveness was not statistically significant (p = .40, .11, .41, .28, respectively). *p*-Uniform's test for publication bias (p = .63, .75, 1.00, .98, .94) as well as the likelihood-ratio test of 3PSM and 4PSM (3PSM: p = .48, .99, .31, .45, .96; 4PSM: p = .72, .65, .13, .70, 1.00; in the order: psychiatric symptoms, target problems, social functioning, personality functioning and overall effectiveness, respectively) yielded no statistically significant results for all outcome domains, suggesting that an adjustment for publication bias was not indicated. However, it is noteworthy that due to the small number of primary studies the bias detection tests have a small power. It can be useful, though, to consider the adjusted estimate even when the bias test is nonsignificant (see ESM 1 for all bias-corrected estimates).

The parameter estimates corrected for publication bias by PET and PEESE are presented in ESM 1. According to Stanley and Doucouliagos's (2014) recommendation on the conditional PET-PEESE estimator, we should use the estimates given by PET because PET is *not* statistically significant for all outcome domains (*p*s > .37). All estimates for the effect size given by PET are smaller than ours and even in the negative range for personality functioning. However, given the inaccurate performance of this approach shown in recent simulation studies (Carter et al., 2017), these estimations should not be taken seriously (especially with small study-level sample sizes and in light of moderate to substantial heterogeneity).

In conclusion, all applied bias detection methods suggested the absence of publication bias. The Egger test was solely statistically significant for the domain personality functioning. This result could be explained by three studies characterized by high standard errors and large effect sizes (Bachar et al., 1999; Bateman & Fonagy, 1999; Gregory et al., 2008). In light of nonsignificant results of *p*-uniform, 3PSM and 4PSM for publication bias, we focus on our main outcomes of the random-effects meta-analysis presented in Table 5. We considered them to be the currently best possible estimates, extensively scrutinized by elaborated statistical tests for publication bias. Given the small number of studies and the low power of bias detection tests, however, this should be seen as a preliminary statement, waiting for updates when more primary studies are available.

Discussion

The empirical evidence for LTPP is still a controversial issue. Recent meta-analyses come to conflicting conclusions about whether LTPP is more efficacious than other forms of therapy. We aimed to reproduce the most recent meta-analysis by Leichsenring et al. (2013) and to conduct an updated meta-analysis which adds three primary studies and applies recent developments in statistical bias detection and correction.

Replication of Leichsenring et al. (2013)

We replicated the direction and general tendency of the meta-analytic results achieved by Leichsenring et al. (2013), but our replicated effect sizes were slightly smaller and showed higher heterogeneity. The largest difference was in the domain psychiatric symptoms, where we found a *g* of 0.28 compared to the originally reported 0.43. One reason for the different findings might be that our categorization of outcome measures slightly differed. Additionally, we assumed that Leichsenring et al. (2013) calculated standard Hedges' *g* for the study by Bressi et al. (2010) instead of ppc-*g* as they did for the other studies. Exceptionally, the single standard Hedges' *g* calculations for the five outcome domains of this primary study were substantially higher than the ppc-*g* calculations (see Table 1 in ESM 1), which could partly explain the higher effects sizes found by Leichsenring et al. (2013). We thank Falk Leichsenring for sending us their interim results and approving to make them public (see ESM 1), but without their raw data of the primary studies, it is not possible to fully explain the difference in the size of the effects. In line with Lakens et al. (2017) we recommend that future meta-analysts should (a) clearly indicate which data were used to calculate an effect size, (b) specify all individual effect sizes and equations that are used to calculate them, (c) explain how multiple effect size estimates from the same primary study are combined, and (d) share raw data acquired from original authors or unpublished research reports. We tried to follow all these recommendations and the ones described in the method section in order to facilitate the reproduction and update of our meta-analysis.

Updated Meta-Analysis

Summary and Discussion of the Level of Evidence

In the second part of our study, the updated meta-analysis, the results were quite similar to our replicated effect sizes, but slightly smaller. For all but one outcome domain, we received small, but statistically significant effect sizes (see Table 5). The small effect size for the outcome domain personality functioning could be labeled as marginally significant, but should be interpreted cautiously because it builds on a narrow data basis and was not robust to our sensitivity analysis. Consequently, the effect size of target problems may as well be regarded as marginally insignificant ($p = .04$). Furthermore, in all outcome domains the effect sizes were qualified by moderate to substantial heterogeneity, suggesting that not a single underlying effect was measured, but that the summary estimate for each domain was an average across measures of multiple effects. This seems plausible because we included different psychiatric disorders, different forms of LTPP, and different forms of control treatments. A possible solution to reduce heterogeneity would have been to conduct separate meta-analyses of subsets, which was not possible at this stage because the data basis was still too narrow for meaningful subgroup analyses. The amount of heterogeneity could be illustrated by our prediction intervals ranging from small to medium effects in favor of the control group to medium to large effects in favor of the LTPP group for all outcome measures. However, our prediction intervals should also be interpreted cautiously because the estimate of the prediction interval is imprecise if it is based on imprecise heterogeneity estimates based on only few studies (IntHout et al., 2016). Furthermore, our assessment of the risk of bias within the primary studies suggested that our set of studies carried an unclear risk of bias, resulting from several methodological shortcomings in the primary studies (see Figure 3). Notably, despite particular difficulties for long-term treatments, the majority of studies applied treatment manuals and checked treatment integrity. It is also noteworthy that the primary studies by Bateman and Fonagy (2009) and Fonagy et al. (2015) applying the most advanced psychoanalytic psychotherapy and research standards revealed a medium overall effect (0.6 and 0.5, respectively) compared to the control group. Nevertheless, in light of moderate to substantial heterogeneity and an unclear risk of bias due to methodological shortcomings in most of the primary studies, our small, statistically significant summary effects should be interpreted cautiously.

To specify the scope of our findings, we concluded to have found small summary effect sizes of LTPP which (1) apply to our pre-defined combination of disorders called complex mental disorders, (2) are common to different forms of LTPP and (3) result from a comparison primarily to other forms of long-term psychotherapy (mean session ratio = 1.67). More specific evidence of the efficacy of LTPP cannot be provided at this time, as the data basis is still too narrow. Our findings were robust to an extensive assessment of publication bias and different ways of calculating Hedges' *g*.

According to the recent APA model by Tolin et al. (2015), a clear recommendation for the empirical support of LTPP can, thus, not be given so far. A classification into the categories *very strong*, *strong*, or *weak* empirical support would at least require meaningful subgroup analyses of specific forms of LTPP (e.g., mentalization based therapy or transference focused psychotherapy) applied to specific disorders (e.g., borderline personality disorder) controlled for by homogeneous control treatments (at best other specialized forms of psychotherapy). Additionally, the quality of evidence needs to be rated according to the GRADE system (Atkins et al., 2004; Guyatt et al., 2008). First, we cannot adhere to the first quality criterion, as we were not able to include a wide range of studies in our analysis due to the narrow data basis. Furthermore, some studies were also characterized by major limitations. Second, our studies do not vary slightly, as required by the GRADE system, but moderately to substantially. Third, our CIs are narrower than those of Smit et al. (2012) and Leichsenring et al. (2013), but still cross the benchmark of 0.2, possibly suggesting a negligible effect. The quality and certainty of our evidence may, therefore, be considered as moderate to low. Furthermore, only five studies measured and reported clinically significant change as required by the model. In total, Tolin et al. (2015) would describe LTPP as still lacking sufficient evidence of efficacy.

However, we still criticize the concept of the APA as not grasping the complexity of mental disorders since the required specificity seems unrealistic, especially in light of a high co-morbidity of mental disorders (Orlinsky, 2008; Seligman, 1995). Hence, on the one hand, we call for an extension of international guidelines to deal with the complexity especially of severe disorders and with the particular challenges of conducting a long-term study (e.g., treatment manuals, fewer studies possible, etc.; Benecke, Huber, Schauenburg, & Staats, 2016). On the other hand, we demand more studies of LTPP oriented on current standards of psychotherapy research such as the LAC depression study (Beutel et al., 2016) and the APD study (Benecke, Huber, Staats, et al., 2016) which are currently being conducted. Given the aforementioned necessity of long-term treatments for patients with complex mental disorders, more funds should be provided for such studies.

Regarding our effect of LTPP as a summary effect of multiple effects in a random effects model, we still consider our findings to have some practical and clinical relevance. There is no agreed upon definition for a clinically relevant effect, though (Steinert, Munder, Rabung, Hoyer, & Leichsenring, 2017). Cuijpers, Turner, Koole, Dijke, and Smit (2014), for example, proposed a threshold of 0.24 for depressive disorders. Leichsenring et al. (2015) recommended 0.5. Since no agreement exists so far, we cautiously assume small clinically relevant effects of LTPP according to Cuijpers et al. (2014). More research is urgently necessary to corroborate these findings. Future research should examine which patients may need long-term psychotherapy and which patients may sufficiently benefit from short-term psychotherapy, irrespective of whether the treatments are rooted in cognitive behavior therapy, psychoanalytic psychotherapy or another bona fide treatment approach (Leichsenring et al., 2013).

Limitations and Future Perspectives

Besides the scarcity of randomized, controlled trials, further limitations of our meta-analysis need to be addressed. Several primary studies suffered from methodological shortcomings (see Figure 3). Hence, our set of studies carried an unclear risk of bias. Especially the issue of a successful blinding of outcome assessment needs to be addressed in future research. In general, future studies on LTPP should more carefully abide by international guidelines and quality criteria (Higgins et al., 2011; Tolin et al., 2015). Additionally, only seven studies reported follow-up data. For a discipline assuming that treatment effects carry on and even continue to improve after the treatment was finished (Huber et al., 2013; Leichsenring & Rabung, 2008; Town et al., 2012), more extensive follow-up data should be provided in the future. Keeping our narrow data basis in mind, we found some evidence that follow-up data may in fact yield higher effect sizes than post-treatment data. Notably, Benecke, Huber, Schauenburg, et al. (2016) reasonably argued that a late follow-up assessment at the same point in time constitutes the most adequate point in time for comparing short-term and long-term treatments. Given our very limited data base of follow-up data, this was not possible in our study. Our post-treatment assessment at the point in time when LTPP was finished might have been a benefit of LTPP because for short term patients, this post-assessment is actually a follow-up assessment. Short-term patients might have already acquired a more realistic view of their life during this follow-up period. However, the majority of control conditions were other forms of long-term treatments with a more similar duration.

Another limitation may be seen in the heterogeneity of control conditions. Five out of 14 control interventions were treatment as usual conditions representing a relatively weak comparator. However, the other nine control conditions were other (specialized) forms of psychotherapy providing stronger comparators. Future research on LTPP should only include strong comparators to allow conclusions about the specific mechanisms of change (Chambless & Hollon, 1998; Tolin et al., 2015) or should consider "treatment as usual" versus a more involved control group as a potential moderator. Different forms of LTPP as well as differences in disorders and outcome assessment measures may be also seen as limitations because they probably may have led to our identified moderate to substantial heterogeneity. Psychotropic medication and other forms of therapy as treatment confounders were found in almost all studies. Especially for severe disorders, pharmacotherapy cannot be excluded. However, we call for a more systematic monitoring of treatment confounders in future research of LTPP. Furthermore, investigator allegiance may distort the results of comparative treatment studies (Munder, Flückiger, Gerger, Wampold, & Barth, 2012). As recommended by Luborsky et al. (1999), our team represented a mix of different allegiances: the main author (CW) had an allegiance to (long-term) psychoanalytic therapy, whereas the second author (FS) had no allegiance. Thus, we attempted to minimize an allegiance effect in our research process. Most of the primary studies, however, have been conducted by proponents of LTPP, and, therefore, the allegiance of trialists may be a rival explanation for the advantage of LTPP. Finally, future primary studies and meta-analyses should include cost-efficiency analyses to provide a solid argumentation for the implementation in health care systems.

Final Considerations

Comparing our updated effect sizes of LTPP to those of Leichsenring et al. (2013), our effect sizes were somewhat smaller than their estimates (0.28 vs. 0.40 for overall effectiveness). We hypothesized to find smaller effect sizes than Leichsenring et al. (2013) after accounting for publication bias. Since we did not need to adjust for publication bias, our difference in effect size is probably explained by our smaller session ratio of 1.67 compared to Leichsenring et al.'s (2013) session ratio of 1.96. They compared LTPP primarily to shorter or less intensive forms of psychotherapy, whereas we compared LTPP primarily to other forms of long-term psychotherapy. A clear classification of our comparison condition into short-term or long-term is not possible, though, because six comparison conditions may be classified as short-term (< 40 sessions) and seven comparison conditions may be classified as long-term (≥ 40 sessions). Gregory et al. (2008) did not report on the mean of the comparison condition. In sum, the small effect sizes we found represent the additional gain of 81.8 of LTPP versus 49.0 sessions of other, primarily long-term, forms of psychotherapy. Here, large differences in effect sizes are not to be expected. We tried to quantify this gain by considering the session ratio as a continuous moderator. The session ratio was only associated with the effect size of personality functioning, but in the unexpected direction. Therefore, the session ratio might not be regarded as a valid moderator, implying that other variables than the number of sessions might account for the additional gain. With an increasing data base, future meta-analysts should compare LTPP only to other forms of long-term psychotherapy by conducting non-inferiority and equivalence analyses (see Steinert et al., 2017, for such analyses of STPP).

Conclusion

We found small statistically significant effect sizes for the outcome domains psychiatric symptoms, target problems, social functioning, and overall effectiveness, when comparing LTPP to other, primarily long-term forms of psychotherapy. The effect size for the outcome domain personality functioning was not significant (*p* = .08). It is noteworthy that large differences in effect sizes are not to be expected since the reported effect sizes represent the additional gain of LTPP versus other forms of primarily long-term psychotherapy. Our effect sizes were robust to an extensive assessment of publication bias and different ways of calculating Hedges' *g*, and according to proposed thresholds, we assume some clinical relevance of our findings. Patients suffering from complex mental disorders seem to benefit slightly more from a treatment with LTPP compared to other, primarily long-term forms of psychotherapy. However, in light of heterogeneous data, an unclear risk of bias across the primary studies, and prediction intervals crossing zero these results should be interpreted cautiously. Further research and improved primary studies are urgently needed to corroborate these results.

Plain Language Summary: A Meta-Analysis of the Efficacy of Long-Term Psychoanalytic Psychotherapy

What Is the Aim of This Review?

The aim of this meta-analysis was to find out whether long-term psychoanalytic psychotherapy is more efficacious than other forms of psychotherapy. The authors collected and analyzed all relevant studies to answer this question and found 14 studies.

Key Messages

Long-term psychoanalytic psychotherapy may be more efficacious than other forms of psychotherapy in the treatment of chronic mental disorders, more than one mental disorder, or a personality disorder.

What Was Studied in the Review?

Adult patients suffering from a chronic mental disorder, more than one mental disorder, or a personality disorder might need more extended treatments. A more extended form of psychotherapy is long-term psychoanalytic psychotherapy. It originated from the theories by Sigmund Freud and is defined as a form of psychotherapy in which the patient and the therapist meet once to twice weekly and the therapy takes place in a sitting position for at least 1 year and 40 sessions. As long-term psychoanalytic psychotherapy causes higher direct financial costs because of a higher number of sessions, its positive effects need to exceed those of less intensive treatments.

What Are the Main Results of the Review?

The authors found 14 relevant studies. These studies compared different forms of long-term psychoanalytic psychotherapy (e.g., mentalization-based therapy) to other forms of psychotherapy. Comparators primarily included other (non-psychoanalytic) long-term therapies, such as dialectical behavior therapy, and some forms of short-term treatments, such as short-term cognitive behavior therapy or basic health support. In each study, patients suffering from a chronic mental disorder, more than one mental disorder, or a personality disorder were randomly assigned to either long-term psychoanalytic psychotherapy or a different form of psychotherapy.

The meta-analysis shows that when patients are treated with long-term psychoanalytic psychotherapy, compared to other, primarily long-term forms of psychotherapy:

1. Patients may show slightly less psychiatric symptoms (low-certainty evidence).
2. Patients may show better capacities to manage social or interpersonal situations (low- to moderate-certainty evidence).
3. Patients may show slightly less problems which are specific to their disorder (e.g., impulse control for borderline patients; low-certainty evidence).
4. We are uncertain whether patients show a higher personality structure (very low-certainty evidence).

The range where the actual effects may be shows that long-term psychoanalytic psychotherapy may lead to a small additional gain, but may also lead to little or no additional gain when compared to other, primarily long-term forms of psychotherapy. Notably, when compared to primarily long-term forms of psychotherapy, large differences are not to be expected.

How Up-to-Date Is This Review?

The authors searched for studies that had been published up to June 11, 2017.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1016-9040/a000385

ESM 1. Supplementary material to the meta-analysis of the efficacy of LTPP

References

Abbass, A. A., Hancock, J. T., Henderson, J., & Kisely, S. (2006). Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database of Systematic Reviews, 4*, 1–50. https://doi.org/10.1002/14651858.CD004687.pub3

Abbass, A. A., Kisely, S. R., Town, J. M., Leichsenring, F., Driessen, E., De Maat, S., … Crowe, E. (2014). Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database of Systematic Reviews, 7*, 1–90. https://doi.org/10.1002/14651858.CD004687.pub4

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839–851. https://doi.org/10.1037/0003-066x.63.9.839

Atkins, D., Eccles, M., Flottorp, S., Guyatt, G. H., Henry, D., Hill, S., . . . The GRADE Working Group. (2004). Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches. *BMC Health Services Research, 4*, 38. https://doi.org/10.1186/1472-6963-4-38

Bachar, E., Latzer, Y., Kreitler, S., Berry, E. M. (1999). Empirical comparison of two psychological therapies: Self psychology and cognitive orientation in the treatment of anorexia and bulimia. *The Journal of Psychotherapy Practice and Research, 8*, 115–128.

Bateman, A., & Fonagy, P. (1999). Effectiveness of partial hospitalization in the treatment of borderline personality disorder: A randomized controlled trial. *American Journal of Psychiatry, 156*, 1563–1569. https://doi.org/10.1176/ajp.156.10.1563

Bateman, A., & Fonagy, P. (2008). 8-year follow-up of patients treated for borderline personality disorder: Mentalization-based treatment versus treatment as usual. *American Journal of Psychiatry, 165*, 631–638. https://doi.org/10.1176/appi.focus.11.2.261

Bateman, A., & Fonagy, P. (2009). Randomized controlled trial of outpatient mentalization-based treatment versus structured clinical management for borderline personality disorder. *American Journal of Psychiatry, 166*, 1355–1364. https://doi.org/10.1176/appi.ajp.2009.09040539

Benecke, C., Huber, D., Schauenburg, H., & Staats, H. (2016). Wie können Langzeittherapien mit kürzeren Behandlungen verglichen werden? Designprobleme und Lösungsvorschläge am Beispiel der APS-Studie [How can long-term therapies be compared with shorter term treatment? Design problems and solution proposals exemplified by the APD study]. *Psychotherapeut, 61*, 476–483. https://doi.org/10.1007/s00278-016-0140-1

Benecke, C., Huber, D., Staats, H., Zimmermann, J., Henkel, M., Deserno, H., . . . Schauenburg, H. (2016). A comparison of psychoanalytic therapy and cognitive behavioral therapy for anxiety (panic/agoraphobia) and personality disorders (APD study): Presentation of the RCT study design. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 62*, 252–269. https://doi.org/10.13109/zptm.2016.62.3.252

Beutel, M. E., Bahrke, U., Fiedler, G., Hautzinger, M., Kallenbach, L., Kaufhold, J., . . . Ernst, M. (2016). LAC-Depressionsstudie [LAC depression study]. *Psychotherapeut, 61*, 468–475. https://doi.org/10.1007/s00278-016-0144-x

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Bressi, C., Porcellana, M., Marinaccio, P. M., Nocito, E. P., & Magri, L. (2010). Short-term psychodynamic psychotherapy versus treatment as usual for depressive and anxiety disorders: A randomized clinical trial of efficacy. *The Journal of Nervous and Mental Disease, 198*, 647–652. https://doi.org/10.1097/nmd.0b013e3181ef3ebb

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General, 144*, 796–815. https://doi.org/10.1037/xge0000083

Carter, E. C., Schönbrodt, F. D., Gervais, W., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science 2*, 115–144. https://doi.org/10.1177/2515245919847196

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7–18. https://doi.org/10.1037/0022-006X.66.1.7

Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2007). Evaluating three treatments for borderline personality disorder: A multiwave study. *American Journal of Psychiatry, 164*, 922–928. https://doi.org/10.1176/ajp.2007.164.6.922

Coburn, K. M., & Vevea, J. L. (2017). *weightr: Estimating weight-function models for publication bias*. R package version 1.1.2. Retrieved from https://CRAN.R-project.org/package=weightr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. Los Angeles, CA: Sage.

Crits-Christoph, P., & Barber, J. P. (2000). Long-term psychotherapy. In C. R. Snyder & R. E. Ingram (Eds.), *Handbook of psychological change: Psychotherapy processes & practices for the 21st century* (pp. 455–473). Hoboken, NJ: Wiley. https://doi.org/10.1016/j.jad.2007.08.005

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences, 28*, 21–30. https://doi.org/10.1017/S2045796018000057

Cuijpers, P., Turner, E. H., Koole, S. L., Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety, 31*, 374–378. https://doi.org/10.1002/da.22249

Cuijpers, P., van Straten, A., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size. *Psychological Medicine, 40*, 211–223. https://doi.org/10.1017/S0033291709006114

Dare, C., Eisler, I., Russell, G., Treasure, J., & Dodge, L. (2001). Psychological therapies for adults with anorexia nervosa. *The British Journal of Psychiatry, 178*, 216–221. https://doi.org/10.1192/bjp.178.3.216

de Maat, S., de Jonghe, F., de Kraker, R., Leichsenring, F., Abbass, A. A., Luyten, P., . . . Dekker, J. (2013). The current state of the empirical evidence for psychoanalysis: A meta-analytic approach. *Harvard Review of Psychiatry, 21*, 107–137. https://doi.org/10.1097/HRP.0b013e318294f5fd

de Maat, S., de Jonghe, F., Schoevers, R., & Dekker, J. (2009). The effectiveness of long-term psychoanalytic therapy: A systematic review of empirical studies. *Harvard Review of Psychiatry, 17*, 1–23. https://doi.org/10.1080/10673220902742476

Derogatis, L. R. (1977). *SCL-90-R: Administration, scoring and procedures manual-I for the revised version*. Baltimore, MD: Clinical Psychometric Research.

Doering, S., Hörz, S., Rentrop, M., Fischer-Kern, M., Schuster, P., Benecke, C., . . . Buchheim, P. (2010). Transference-focused psychotherapy v. treatment by community psychotherapists for borderline personality disorder: Randomised controlled trial. *The British Journal of Psychiatry, 196*, 389–395. https://doi.org/10.1192/bjp.bp.109.070177

Eagan, B., Rogers, B., Serlin, R., Ruis, A., Arastoopour Irgens, G., & Williamson Shaffer, D. (2017). *Can we rely on IRR? Testing the assumptions of inter-rater reliability*. Retrieved from https://repository.isls.org/handle/1/275

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891–904. https://doi.org/10.1007/s11192-011-0494-7

Fonagy, P., Rost, F., Carlyle, J.-A., McPherson, S., Thomas, R., Pasco Fearon, R., . . . Taylor, D. (2015). Pragmatic randomized controlled trial of long-term psychoanalytic psychotherapy for treatment-resistant depression: The Tavistock Adult Depression Study (TADS). *World Psychiatry, 14*, 312–321. https://doi.org/10.1002/wps.20267

Fonagy, P., Steele, M., Steele, H., & Target, M. (1998). *Reflexive-function manual: Version 5.0 for application to the adult attachment interview* Unpublished manual, University College, London, UK.

Gabbard, G. O. (2017). *Long-term psychodynamic psychotherapy: A basic text*. Arlington, VA: American Psychiatric Association.

Giesen-Bloo, J., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., van Asselt, T., . . . Arntz, A. (2006). Outpatient psychotherapy for borderline personality disorder: Randomized trial of schema-focused therapy vs transference-focused psychotherapy. *Archives of General Psychiatry, 63*, 649–658. https://doi.org/10.1001/archpsyc.63.6.649

Gregory, R. J., Chlebowski, S., Kang, D., Remen, A. L., Soderberg, M. G., Stepkovitch, J., & Virk, S. (2008). A controlled trial of psychodynamic psychotherapy for co-occurring borderline personality disorder and alcohol use disorder. *Psychotherapy: Theory, Research, Practice, Training, 45*, 28. https://doi.org/10.1037/0033-3204.45.1.28

Gregory, R. J., DeLucia-Deranja, E., & Mogle, J. A. (2010). Dynamic deconstructive psychotherapy versus optimized community care for borderline personality disorder co-occurring with alcohol use disorders: A 30-month follow-up. *The Journal of Nervous and Mental Disease, 198*, 292–298. https://doi.org/10.1097/NMD.0b013e3181d6172d

Grünbaum, A. (1988). *Die Grundlagen der Psychoanalyse: Eine philosophische Kritik* [The foundations of psychoanalysis: A philosophical critique]. Stuttgart, Germany: Reclam.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schunemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal, 336*, 924–926. https://doi.org/10.1136/bmj.39489.470347.AD

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61–85. https://doi.org/10.3102/10769986009001061

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., . . . Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343*, d5928. https://doi.org/10.1136/bmj.d5928

Higgins, J. P. T. Green, S. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley-Blackwell.

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *British Medical Journal, 319*, 670–674. https://doi.org/10.1136/bmj.319.7211.670

Hollon, S. D., & Ponniah, K. (2010). A review of empirically supported psychological therapies for mood disorders in adults. *Depression and Anxiety, 27*, 891–932. https://doi.org/10.1002/da.20741

Huber, D., Henrich, G., Clarkin, J. F., & Klug, G. (2013). Psychoanalytic versus psychodynamic therapy for depression: A three-year follow-up study. *Psychiatry: Interpersonal & Biological Processes, 76*, 132–149. https://doi.org/10.1521/psyc.2013.76.2.132

Huber, D., Henrich, G., Gastner, J., & Klug, G. (2012). Must all have prizes? The Munich psychotherapy study. In R. A. Levy, J. S. Ablon, & H. Kächele (Eds.), *Psychodynamic Psychotherapy Research* (pp. 51–69). New York, NY: Humana Press. https://doi.org/10.1007/978-1-60761-792-1_4

Huber, D., Zimmermann, J., Henrich, G., & Klug, G. (2012). Comparison of cognitive-behaviour therapy with psychoanalytic and psychodynamic therapy for depressed patients – A three-year follow-up study. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 58*, 299–316. https://doi.org/10.13109/zptm.2012.58.3.299

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open, 6*, e010247. https://doi.org/10.1136/bmjopen-2015-010247

Jørgensen, C. R., Bøye, R., Andersen, D., Døssing Blaabjerg, A. H., Freund, C., Jordet, H., & Kjølbye, M. (2014). Eighteen months post-treatment naturalistic follow-up study of mentalization-based therapy and supportive group treatment of borderline personality disorder: Clinical outcomes and functioning. *Nordic Psychology, 66*, 254–273. https://doi.org/10.1080/19012276.2014.963649

Jørgensen, C. R., Freund, C., Bøye, R., Jordet, H., Andersen, D., & Kjølbye, M. (2013). Outcome of mentalization-based and supportive psychotherapy in patients with borderline personality disorder: A randomized trial. *Acta Psychiatrica Scandinavica, 127*, 305–317. https://doi.org/10.1111/j.1600-0447.2012.01923.x

Klar, F. J. (2005). Wirksamkeit individualpsychologisch-psychoanalytischer Psychotherapie [The efficacy of individual psychological-psychoanalytical psychotherapy]. *Zeitschrift für Individualpsychologie, 30*, 28–50.

Knekt, P., Heinonen, E., Härkäpää, K., Järvikoski, A., Virtala, E., Rissanen, J., & Lindfors, O. (2015). Randomized trial on the effectiveness of long- and short-term psychotherapy on psychosocial functioning and quality of life during a 5-year follow-up. *Psychiatry Research, 229*, 381–388. https://doi.org/10.1016/j.psychres.2015.05.113

Knekt, P., Lindfors, O., Härkänen, T., Välikoski, M., Virtala, E., Laaksonen, M. A., . . . Renlund, C. (2008). Randomized trial on the effectiveness of long- and short-term psychodynamic psychotherapy and solution-focused therapy on psychiatric symptoms during a 3-year follow-up. *Psychological Medicine, 38*, 689–703. https://doi.org/10.1017/s003329170700164x

Knekt, P., Lindfors, O., Laaksonen, M. A., Raitasalo, R., Haaramo, P., & Järvikoski, A. (2008). Effectiveness of short-term and long-term psychotherapy on work ability and functional capacity – A randomized clinical trial on depressive and anxiety disorders. *Journal of Affective Disorders, 107*, 95–106. https://doi.org/10.1016/j.jad.2007.08.005

Knekt, P., Virtala, E., Härkänen, T., Vaarama, M., Lehtonen, J., & Lindfors, O. (2016). The outcome of short- and long-term psychotherapy 10 years after start of treatment. *Psychological Medicine, 46*, 1175–1188. https://doi.org/10.1017/s0033291715002718

Kopta, S. M., Howard, K. I., Lowry, J. L., & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology, 62*, 1009–1016. https://doi.org/10.1037/0022-006X.62.5.1009

Korner, A., Gerull, F., Meares, R., & Stevenson, J. (2006). Borderline personality disorder treated with the conversational model: A replication study. *Comprehensive Psychiatry, 47*, 406–411. https://doi.org/10.1016/j.comppsych.2006.01.003

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*, 1–10. https://doi.org/10.1186/s40359-016-0126-3

Lakens, D., LeBel, E. P., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., & Hertogs, R. (2017). *Examining the reproducibility of meta-analyses in psychology: A preliminary report*. Retrieved from https://osf.io/q23ye/

Leichsenring, F., Abbass, A. A., Luyten, P., Hilsenroth, M., & Rabung, S. (2013). The emerging evidence for long-term psychodynamic therapy. *Psychodynamic Psychiatry, 41*, 361–384. https://doi.org/10.1521/pdps.2013.41.3.361

Leichsenring, F., Leweke, F., Klein, S., & Steinert, C. (2015). The empirical status of psychodynamic psychotherapy – An update: Bambi's alive and kicking. *Psychotherapy and Psychosomatics, 84*, 129–148. https://doi.org/10.1159/000376584

Leichsenring, F., & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. *Journal of the American Medical Association, 300*, 1551–1565. https://doi.org/10.1001/jama.300.13.1551

Leichsenring, F., & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: Update of a meta-analysis. *The British Journal of Psychiatry, 199*, 15–22. https://doi.org/10.1176/appi.focus.12.3.336

Leichsenring, F., Rabung, S., & Leibing, E. (2004). The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: A meta-analysis. *Archives of General Psychiatry, 61*, 1208–1216. https://doi.org/10.1001/archpsyc.61.12.1208

Levy, K. N., Clarkin, J. F., Yeomans, F. E., Scott, L. N., Wasserman, R. H., & Kernberg, O. F. (2006). The mechanisms of change in the treatment of borderline personality disorder with transference focused psychotherapy. *Journal of Clinical Psychology, 62*, 481–501. https://doi.org/10.1002/jclp.20239

Lindfors, O., Knekt, P., Heinonen, E., Härkänen, T., & Virtala, E. (2015). The effectiveness of short- and long-term psychotherapy on personality functioning during a 5-year follow-up. *Journal of Affective Disorders, 173*, 31–38. https://doi.org/10.1016/j.jad.2014.10.039

Lindfors, O., Knekt, P., Virtala, E., & Laaksonen, M. A. (2012). The effectiveness of solution-focused therapy and short- and long-term psychodynamic psychotherapy on self-concept during a 3-year follow-up. *The Journal of Nervous and Mental Disease, 200*, 946–953. https://doi.org/10.1097/NMD.0b013e3182718c6b

Linehan, M. M., Comtois, K. A., Murray, A. M., Brown, M. Z., Gallop, R. J., Heard, H. L., ... Lindenboim, N. (2006). Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of General Psychiatry, 63*, 757–766. https://doi.org/10.1001/archpsyc.63.7.757

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., ... Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6*, 95–106. https://doi.org/10.1093/clipsy.6.1.95

Masling J. (Ed.). (1983). *Empirical studies of psychoanalytical theories*. Hillsdale, NJ: Erlbaum.

McMain, S. F., Links, P. S., Gnam, W. H., Guimond, T., Cardish, R. J., Korman, L., & Streiner, D. L. (2009). A randomized trial of dialectical behavior therapy versus general psychiatric management for borderline personality disorder. *American Journal of Psychiatry, 166*, 1365–1374. https://doi.org/10.1176/appi.ajp.2009.09010039

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*, 730–749. https://doi.org/10.1177/1745691616662243

Mertens, W. (2013). Psychoanalyse als Methode, Theorie und Praxis [Psychoanalysis as a methodology, theory and practice]. In W. Mertens, C. Benecke, L. Gast, & M. Leuzinger-Bohleber (Eds.), *Psychoanalyse im 21. Jahrhundert: Eine Standortbestimmung* (pp. 13–32). Stuttgart, Germany: Kohlhammer.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364–386. https://doi.org/10.1177/1094428106291059

Munder, T., Flückiger, C., Gerger, H., Wampold, B. E., & Barth, J. (2012). Is the allegiance effect an epiphenomenon of true efficacy differences between treatments? A meta-analysis. *Journal of Counseling Psychology, 59*, 631–637. https://doi.org/10.1037/a0029571

Munder, T., Flückiger, C., Leichsenring, F., Abbass, A. A., Hilsenroth, M. J., Luyten, P., & Wampold, B. E. (2019). Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiology and Psychiatric Sciences, 28*(3), 268–274. https://doi.org/10.1017/S2045796018000355

Orlinsky, D. (2008). Die nächsten 10 Jahre Psychotherapieforschung: Eine Kritik des herrschenden Forschungsparadigmas mit Korrekturvorschlägen [The next 10 years of psychotherapy research: A critique of the prevailing research paradigm]. *Psychotherapie Psychosomatik Medizinische Psychologie, 58*, 345–354. https://doi.org/10.1055/s-2008-1067444

Popper, K. R. (1972). *Conjectures and refutations: The growth of scientific knowledge*. London, UK: Routledge and Kegan Paul.

Poulsen, S., Lunn, S., Daniel, S. I., Folke, S., Mathiesen, B. B., Katznelson, H., & Fairburn, C. G. (2014). A randomized controlled trial of psychoanalytic psychotherapy or cognitive-behavioral therapy for bulimia nervosa. *American Journal of Psychiatry, 171*, 109–116. https://doi.org/10.1176/appi.focus.120410

Puschner, B., Kraft, S., Kächele, H., & Kordy, H. (2007). Course of improvement over 2 years in psychoanalytic and psychodynamic outpatient psychotherapy. *Psychology and Psychotherapy: Theory, Research and Practice, 80*, 51–68. https://doi.org/10.1348/147608306x107593

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Seligman, M. E. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist, 50*, 965–974. https://doi.org/10.1037/0003-066x.50.12.965

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666–681. https://doi.org/10.1177/1745691614553988

Smit, Y., Huibers, M. J., Ioannidis, J. P., van Dyck, R., van Tilburg, W., & Arntz, A. (2012). The effectiveness of long-term psychoanalytic psychotherapy – A meta-analysis of randomized controlled trials. *Clinical Psychology Review, 32*, 81–92. https://doi.org/10.1016/j.cpr.2011.11.003

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60–78. https://doi.org/10.1002/jrsm.1095

Steinert, C., Munder, T., Rabung, S., Hoyer, J., & Leichsenring, F. (2017). Psychodynamic therapy: As efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *American Journal of Psychiatry, 174*, 943–953. https://doi.org/10.1176/appi.ajp.2017.17010057

Svartberg, M., Stiles, T. C., & Seltzer, M. H. (2004). Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for cluster C personality disorders. *American Journal of Psychiatry, 161*, 810–817. https://doi.org/10.1176/appi.ajp.161.5.810

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice, 22*, 317–338. https://doi.org/10.1111/cpsp.12122

Town, J. M., Diener, M. J., Abbass, A. A., Leichsenring, F., Driessen, E., & Rabung, S. (2012). A meta-analysis of psychodynamic psychotherapy outcomes: Evaluating the effects of research-specific procedures. *Psychotherapy, 49*, 276–290. https://doi.org/10.1037/a0029564

van Aert, R. C. M. (2017). *p*-uniform. Retrieved from https://github.com/RobbievanAert/puniform van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on *p* values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science, 11*, 713–729. https://doi.org/10.1177/1745691616650874

van Asselt, A. D., Dirksen, C. D., Arntz, A., Giesen-Bloo, J. H., van Dyck, R., Spinhoven, P., . . . Severens, J. L. (2008). Out-patient psychotherapy for borderline personality disorder: Cost-effectiveness of schema-focused therapy v. transference-focused psychotherapy. *The British Journal of Psychiatry, 192*, 450–457. https://doi.org/10.1192/bjp.bp.106.033597

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293–309. https://doi.org/10.1037/met0000025

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33*, 1111–1115. https://doi.org/10.1001/archpsyc.1976.01770090101010

Werner, C., & Langenmayr, A. (2006). *Die Bedeutung der frühen Kindheit. Psychoanalyse und Empirie* [The significance of early childhood. Psychoanalysis and Empirical Evidence]. Göttingen, Germany: Vandenhoeck & Ruprecht.

Zimmermann, J., Löffler-Stastka, H., Huber, D., Klug, G., Alhabbo, S., Bock, A., & Benecke, C. (2015). Is it all about the higher dose? Why psychoanalytic therapy is an effective treatment for major depression. *Clinical Psychology & Psychotherapy, 22*, 469–487. https://doi.org/10.1002/cpp.1917

Zipfel, S., Wild, B., Groß, G., Friederich, H.-C., Teufel, M., Schellberg, D., . . . Herpertz, S. (2014). Focal psychodynamic therapy, cognitive behaviour therapy, and optimised treatment as usual in outpatients with anorexia nervosa (ANTOP study): Randomised controlled trial. *The Lancet, 383*, 127–137. https://doi.org/10.1016/s0140-6736(13)61746-8

History
Received March 16, 2018
Revision received July 6, 2019
Accepted August 6, 2019
Published online December 6, 2019

Acknowledgments
We thank Falk Leichsenring and all other authors who provided us with additional information and/or data. Furthermore, we are grateful to Wiebke Goltz, Martina Stumpp, and Jens Christian for their comments on an earlier version of the manuscript. We also thank Larissa Sust for double coding our raw data and outcome measures.

Open Data
We embrace the values of openness and transparency in science (http://www.researchtransparency.org/). The preregistration, open data, and reproducible scripts for all data analyses reported in this paper can be accessed at https://osf.io/vec5d/.

Christian Franz Josef Woll
Department of Psychology
Clinical Psychology of Children and Adolescents and Psychology of Interventions
Ludwig-Maximilians-Universität Munich
Leopoldstr. 13
80802 Munich
Germany
christian.woll@psy.lmu.de

Christian Franz Josef Woll (MSc, Clinical Psychology) is a research fellow at the Department of Clinical Psychology of Children and Adolescents at the Ludwig-Maximilians-Universität Munich. His major research interests include the impact of psychiatric disorders on caregiver-infant interaction, meta-analytic methods, and current developments in open science. He also is a candidate of psychoanalytic training at the Akademie für Psychoanalyse und Psychotherapie, Munich.

Felix Schönbrodt (PhD) is a principle investigator at the Ludwig-Maximilians- Universität Munich and the managing director of the LMU Open Science Center. He obtained his PhD in psychology in 2010 at the Humboldt-University Berlin and received his habilitation 2014 at the Ludwig-Maximilians-Universität Munich. His research interests include implicit and explicit motives, quantitative methods in Bayesian statistics and meta-analysis, data visualization, and issues revolving around open science and the replicability of research.