Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research

F. Leichsenring14†, A. Abbass2, M. J. Hilsenroth3, F. Leweke1, P. Luyten4,5, J. R. Keefe6, N. Midgley7,8, S. Rabung9,10, S. Salzer11,12 and C. Steiner1

Replicability of findings is an essential prerequisite of research. For both basic and clinical research, however, low replicability of findings has recently been reported. Replicability may be affected by research biases not sufficiently controlled for by the existing research standards. Several biases such as researcher allegiance or selective reporting are well-known for affecting results. For psychotherapy and pharmacotherapy research, specific additional biases may affect outcome (e.g. therapist allegiance, therapist effects or impairments in treatment implementation). For meta-analyses further specific biases are relevant. In psychotherapy and pharmacotherapy research these biases have not yet been systematically discussed in the context of replicability. Using a list of 13 biases as a starting point, we discuss each bias’s impact on replicability. We illustrate each bias by selective findings of recent research, showing that (1) several biases are not yet sufficiently controlled for by the presently applied research standards, (2) these biases have a pernicious effect on replicability. For the sake of research credibility, it is critical to avoid these biases in future research. To control for biases and to improve replicability, we propose to systematically implement several measures in psychotherapy and pharmacotherapy research, such as adversarial collaboration (inviting academic rivals to collaborate), reviewing study design prior to knowing the results, triple-blind data analysis (including subjects, investigators and data managers/statisticians), data analysis by other research teams (crowdsourcing), and, last not least, updating reporting standards such as CONSORT or the Template for Intervention Description and Replication (TIDieR).

Received 15 December 2015; Revised 14 November 2016; Accepted 15 November 2016

Key words: Efficacy, evidence-based medicine, psychotherapy research, replicability, risk factors.

Introduction

Replicability of findings is an essential prerequisite of research (Popper, 1959, p. 45). It can be defined as obtaining the same finding with other (random) samples representative of individuals, situations, operationalizations, and time points for the hypothesis tested in the original study (Brunswik, 1955; Asendorpf et al. 2016). It is a prerequisite for valid conclusions (Asendorpf et al. 2016). However, results that are replicable are not necessarily valid. This is true, for example, if they are based on the same errors in measurement.

For cognitive and social-personality psychology, recent research showed that depending on the criterion used, only 36–47% of the original studies were successfully replicated (Open Science Collaboration, 2015). This result led some authors to the conclusion that there is a ‘replication crisis’ in psychological science (Carey, 2015). There is evidence suggesting similar problems for many areas of clinical research (Ioannidis, 2005a, 2009; Nuzzo, 2015; Tajika et al. 2015). For psychotherapy and pharmacotherapy a recent study reported low rates of replication (Tajika et al. 2015). Low replicability of clinical research is even more alarming since results that are neither replicable nor valid may lead to questionable treatment recommendations, may promote suboptimal clinical outcomes, and may influence decisions of insurance companies, policy makers, and funding organizations.

For improving replicability in psychotherapy and pharmacotherapy research, identification of risk factors...
for non-replicability is important. Biases in research are well-known for affecting results (e.g., Ioannidis, 2005). In this article, we discuss several research biases with regard to their effect on replicability. Finally, we suggest measures to control for these risk factors and to improve replicability of psychotherapy and pharmacotherapy research.

Method

Bias can be defined as ‘the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced’ (Ioannidis, 2005, p. 697). We used a list of well-known biases made up by Ioannidis (2005) as a starting point (e.g., researcher allegiance, selective reporting, small studies, or small effects sizes) which we complemented by biases specific to psychotherapy and pharmacotherapy research such as impairments in treatment integrity, therapist or supervisor allegiance, therapist/clinician effects (e.g., Wampold & Imel, 2015). In addition we addressed specific biases relevant to meta-analyses in the field. In total, we examined thirteen biases presented in Table 1. For psychotherapy and pharmacotherapy research these biases have not yet been systematically discussed in the context of replicability. We illustrate each bias by selective findings of recent research. We did not aim at examining a random sample of studies, but rather chose to highlight the relevance of these risk factors by demonstrative examples.

Results

Allegiance effects

Researcher allegiance

In biomedical research, conflicts of interest and prejudice are common, but only sparsely reported, let alone controlled for (Ioannidis, 2005; Dragioti et al. 2015). In psychotherapy research, researcher’s own allegiances may have been found to heavily influence the results of comparative studies in psychotherapy (Luborsky et al. 1999). No less than 69% of variance in outcomes in psychotherapy research were found to be explained by the researchers allegiances, which was therefore called a ‘wild card’ in comparative outcome research. As recent studies have corroborated these earlier findings (Munder et al. 2012; Falkenström et al. 2013), still today researcher allegiance is a widely uncontrolled ‘wild card’ in research. Researcher allegiances are difficult to control for as they often operate on an implicit or unconscious level and are not necessarily the result of deliberate attempts to distort results (Nuzzo, 2015). They often find expression in design features such as the selection of outcome measures (Munder et al. 2011), poor implementation of unfavored treatments (Munder et al. 2011) or uncontrolled therapist allegiance (Falkenström et al. 2013). As there is no statistical algorithm to assess bias, human judgment is required to detect such effects (Higgins et al. 2011).

It is of note that allegiance per se does not necessarily affect replicability. This is only the case if allegiances are not balanced between the study conditions. Allegiances may be balanced, for example, by including researchers, therapists and supervisors with each of whom being alleged to (only) one of the treatments compared (‘adversarial collaboration’, Mellers et al. 2001). Alternatively, treatment studies may be carried out by researchers who are not alleged to either of the treatments under study (Wampold & Imel, 2015). This was the case, for example in the randomized controlled trial (RCT) by Elkin et al. (1989) comparing cognitive-behavioral therapy (CBT), interpersonal therapy (IPT) and pharmacotherapy in the treatment of depression.

A recent RCT may serve as an example for an uncontrolled allegiance effect. In this study cognitive therapy (CT) and ‘Rogerian supportive therapy’ (RST) were compared in borderline personality disorder (Cottraux et al. 2009). Several features of the design, the data analysis and the presentation of results suggest allegiance effects, both in researchers and therapists. (1) For CT the therapists received three 2-day workshops, whereas the training in RST encompassed only 10 h of role-play. (2) The training in CT was carried out by a specialist, but it is not clear by whom the training in RST was conducted. (3) The treatments in both groups were carried out by the same therapists who had a CBT diploma, raising the question of therapist allegiance (see ‘Therapist allegiance’ section below), which may be additionally fostered by the differences in training duration. (4) No significant differences between the treatments were found in the primary outcome (response) at any time of measurement (Cottraux et al. 2009). The authors used several secondary outcome measures and carried out a large number of significance tests, 13 for each the three times of assessment, without, however, any adjustment for type-I error. In only six of these 39 tests, was a statistically significant difference in outcome in favor of CT found. It is not known how many of them are due to chance. (5) Thus, the majority of results suggest that no differences in outcome between CT and RST exist, especially in the primary outcome. The authors, however, concluded (Cottraux et al. 2009, p. 307): ‘CT … showed earlier positive effects on hopelessness and impulsivity, and demonstrated better...
long-term outcomes on global measures of improvement. Thus, from a large number of non-significant differences, the authors picked out the few differences in favor of CT (selective interpretation) of which some may also be due to chance. Taken together, the issues listed above raise the question of a researcher and therapist allegiance in favor of CT. These biases may affect replicability: In more balanced comparisons the results may not be replicated.

**Therapist allegiance**

If the same therapists perform the different treatments being compared, a therapist bias may be introduced in the design, especially if therapists show a specific therapeutic orientation. This was the case, for example in the RCT by Cottraux et al. (2009) discussed above. In pharmacotherapy, the effects of the psychiatrist may be larger than the medication effects (McKay et al. 2006; Wampold & Imel, 2015, p. 170). These results suggest that therapist allegiance may play an important role in pharmacotherapy as well.

**Supervisor allegiance**

A comparable effect may result if the treatments being compared are supervised by the same supervisor (Table 1).

Due to space limitations, we can only present selected examples for each bias. Further examples for

<table>
<thead>
<tr>
<th>Risk factors</th>
<th>Proposed measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Allegiances</td>
<td>Triple-blind data analysis (including subjects, investigators and data managers/statisticians); data analysis by other research teams (crowdsourcing), adversarial collaboration (inviting academic rivals to collaborate); including pertinent items on researcher allegiance in guidelines</td>
</tr>
<tr>
<td>1.1 Researcher allegiance</td>
<td>Treatments are carried out by experts of the respective approach, treatments in the different conditions are not carried out by the same therapists. The same applies to treatment supervisors</td>
</tr>
<tr>
<td>1.2 Therapist allegiance</td>
<td>Therapists are supervised by experts in the respective approach. No supervision of different treatments by the same therapist</td>
</tr>
<tr>
<td>1.3 Supervisor allegiance</td>
<td>Blinded reviewers; review of study design prior to knowing the results; no anonymous reviews; public control of reviewer decisions (especially for grant applications)</td>
</tr>
<tr>
<td>1.4 Reviewer allegiance</td>
<td>Editor allegiance may be reduced by measures for a more open and transparent journal policy, e.g. registered reports</td>
</tr>
<tr>
<td>1.5 Editor allegiance/policy</td>
<td>Including researchers of the rival approaches; including items in reporting guidelines addressing structural equivalence of treatments (e.g. selection and training of therapists, supervision, duration of treatments, adherence measurement)</td>
</tr>
<tr>
<td>2. Impaired treatment integrity (‘strawman’ therapies)</td>
<td>Including researchers of the rival approaches; including items in reporting guidelines addressing structural equivalence of treatments (e.g. selection and training of therapists, supervision, duration of treatments, adherence measurement)</td>
</tr>
<tr>
<td>3. Ignoring therapist effects</td>
<td>Taking therapist effects in data analysis systematically into account; report of effect sizes for therapist effects (ICC)</td>
</tr>
<tr>
<td>4. Small effect sizes: overemphasizing small differences</td>
<td>Differentiating between statistically and clinically significant findings; a priori defining a clinically meaningful threshold in upfront trial registration</td>
</tr>
<tr>
<td>5. Flexibility in design: multiple outcome measures and selective outcome reporting</td>
<td>Upfront study registration including primary and secondary outcomes; focus on ITT analyses</td>
</tr>
<tr>
<td>6. Small sample sizes</td>
<td>Performing higher powered studies when addressing relatively established findings; meta-analyses achieve higher power than small individual studies</td>
</tr>
<tr>
<td>7. Publication bias</td>
<td>Upfront trial registration; increased publication of non-significant results (change in editor policy), acceptance of manuscripts before results are known</td>
</tr>
<tr>
<td>8.1 Selective inclusion of non-bona fide studies in meta-analyses</td>
<td>Upfront registration; measures described above to control for allegiances of researchers, reviewers and editors</td>
</tr>
<tr>
<td>8.2 Selective exclusion of bona fide studies in meta-analyses</td>
<td>Upfront registration, measures described above to control for allegiances of researchers, reviewers and editors</td>
</tr>
</tbody>
</table>
researcher, therapist and/or supervisor allegiance were discussed, for example, by Wampold & Imel (2015, pp. 120–128). Measures to control for allegiance effects are proposed below (see Conclusions and Table 1).

Reviewer allegiance – a dark field in research

Within the peer review system, researchers also serve as reviewers for journals or grant applications. Thus, allegiances in reviewers may be present as well. They may lead to unbalanced decisions about rejection or acceptance of manuscripts or grant applications, distorting the available evidence and affecting its replicability. Whereas there is substantial evidence for the researcher allegiance effect, research on reviewer allegiance is essentially non-existent – it is a dark field in research. Experimental studies, however, suggest that reviewers tend to accept results that are consistent with their expectations, but tend to question the study if this is not the case (Fugelsang et al. 2004). According to a recent study, 83% of researchers in Germany doubt that reviewers are impartial (Spiwak, 2016). As another problem, recommendations given in review articles were found to seriously deviate from available evidence, possibly suggesting reviewer allegiances (Antman et al. 1992; Ioannidis, 2005).

Journal editors’ allegiance and publication policy

Whereas publication bias is well-known (Rothstein et al. 2005), journal editors’ allegiances are another dark field of research, with no data available. As other researchers, editors may be biased as well. If submitted articles are rejected because the results are not consistent with the journal’s editorial policy (‘editor allegiance’), a publication bias may result that can be expected to affect replicability. For the credibility of research, a more open journal policy is required (Nuzzo, 2015).

Impaired treatment integrity: ‘strawman’ therapies

Treatment integrity is defined as the degree to which treatments are carried out as originally intended (Yeaton & Sechrest, 1981; Kazdin, 1994). This definition applies to pharmacotherapy research as well. If the pharmacological treatment is described in a treatment manual with regard to dose, treatment duration and clinical management (e.g. Elkin et al. 1985; Davidson et al. 2004, p. 1006), also the pharmacological treatment may be implemented more or less consistent with the manual and the study protocol. As psychiatrist effects may have a stronger impact on outcome than the medication (McKay et al. 2006; Wampold & Imel, 2015, p. 170), they may play an important part for therapy integrity.

Despite the importance of therapy integrity, a review reported that in more than 96% of RCTs published in the most influential psychiatric and psychological journals the quality of treatment integrity procedures was low (Perepletchikova et al. 2007).

Treatment integrity implies that for each treatment a valid version of the treatment is adequately implemented. Already in one of the earliest meta-analyses within the field, however, Smith et al. (1980, p. 119) reported that often the comparison condition was implemented as a ‘strawman’ condition intended to fail. In contrast, bona fide therapies are (a) delivered by trained therapists, (b) offered to the therapeutic community as viable treatments (e.g. based on professional books or manuals), and (c) contain specific treatment components based on theories of change (Wampold et al. 1997). If a non-bona fide treatment is implemented as a comparator, treatment effects may be overestimated and not replicable.

As an additional problem, a treatment may be implemented as intended – without being a bona fide therapy. This is the case if in the conceptualization of a method of, for example, CBT, psychodynamic therapy (PDT) or interpersonal therapy included in the study protocol essential treatment elements are omitted (neutering of treatment). As a consequence, the treatment may be implemented in accordance with the study protocol and the study may be described and reported in accordance with recent guidelines such as the Consolidated Standards of Reporting Trials (CONSORT, Moher et al. 2010) or the Template for Intervention Description and Replication (TIDieR; Hoffmann et al. 2014). The problem in treatment integrity will not come to the fore. In this case, demonstrated treatment integrity is orthogonal from ‘intent-to-fail’ treatments.

An RCT comparing PDT to CBT in adolescents with post-traumatic stress disorder (PTSD) may serve as an example (Gilboa-Schechtman et al. 2010). Several design features suggest imbalances in treatment implementation. (1) In the PDT condition, the therapists were trained for 2 days, whereas the CBT therapists were trained for 5 days. (2) Therapists in the CBT condition were trained by Edna Foa, a world expert in PTSD, whereas the therapists in PDT were trained by one of the study authors (L.R.), whose expertise in PDT is not clear. (3) Maybe most importantly, therapists in PDT were not allowed to directly address the trauma, but instead were requested to focus on an ‘unresolved conflict’ (e.g. dependence-independence, or passivity-activity) (Gilboa-Schechtman et al. 2010, p. 1035), a psychological constellation obviously not primarily relevant to the trauma-induced psychopathology. Thus, therapists were instructed to avoid addressing an issue that was highly relevant to.
patients who entered treatment for their PTSD symptoms. This is especially perplexing, since existing methods of PDT for PTSD explicitly include a focus on the trauma (Horowitz & Kaltreider, 1979; Woeller et al. 2012). Thus, therapists were instructed to ignore primary aspects of their treatment model.

The study by Gilboa-Schechtman et al. (2010) highlights the problem noted above: If a neutered version of an originally bona fide treatment is included in the study protocol, the treatment may be implemented as intended – without being a bona fide therapy, a problem presently not detected by standards such as TIDieR.

Neutering, however, may not only refer to specific, but also to non-specific treatment components.

In an RCT by Snyder & Wills (1989) behavioral and insight-oriented marital therapy were equally effective posttherapy, but significantly less couples of the insight-oriented therapy group were divorced in the 4-year follow-up (Snyder et al. 1991). As emphasized by Jacobson (1991), however, non-specific interventions were included in the insight-oriented treatment manual, but not in the behavioral manual, introducing an advantage for insight-oriented therapy.

Furthermore, not only active treatments may be neutered, but also placebo controls. This effect was demonstrated in an earlier meta-analysis by Dush et al. (1983) for several studies on Meichenbaum’s method of self-statement modification which yielded considerably lower effects for placebos (and larger effects for Meichenbaum’s method) when studies were carried out by Meichenbaum himself.

Further examples for neutering comparison conditions were presented by Wampold & Imel (2015, p. 120–128) who critically discussed the studies by Clark et al. (1994) or Foa et al. (1991). Thus, neutering of comparison conditions is not uncommon, showing that the examples we are presenting do not represent arbitrarily selected rare events.

In sum, impairing treatment integrity may lead to results that are neither replicable nor valid. Especially the recent studies discussed above illustrate that the presently existing standards such as CONSORT or TIDierR do not yet prevent impairments in treatment implementation. Updating research standards specifically for this problem is required.

Ignoring therapist effects

Clinicians vary in their efficacy, both within and between treatment conditions, not only in psychotherapy, but also when delivering pharmacotherapy (McKay et al. 2006; Wampold & Imel, 2015, p. 170). As a consequence, observations are not independent, such as the outcomes of patients X and Y treated by the same therapist Z (Wampold & Imel, 2015). For this reason, therapists need to be statistically taken into account as a nested random factor (Wampold & Imel, 2015), although larger sample sizes are needed to achieve this (Wampold & Imel, 2015). Failure to do so may result in increased type I errors and overestimating treatment effects (Wampold & Imel, 2015, p. 164). Thus, ignoring therapist effects may lead to false conclusions about treatment efficacy and to results that are not replicable (e.g. ‘treatment A is superior to B’). Estimates for the reduction of significant differences between treatments depending on the size of therapist effects and the number of patients treated per therapist were recently provided by a simulation study (Owen et al. 2015). With small, medium and large effect sizes for therapist effects (ICC = 0.05, 0.10, 0.20), for example, only 80%, 65% and 35% of simulated significant differences were still significant after adjusting for therapist effects, assuming that on average 15 patients are treated per therapist (Owen et al. 2015). With more patients per therapist, the reduction is even larger (Owen et al. 2015). Because many trials are underpowered to detect therapist effects, even though therapist effects are not statistically significant, the pernicious effects on error rates and effect sizes are present and these problems are exacerbated when there are fewer therapists (Wampold & Imel, 2015). Increasing the risk for type I error and overestimating treatment effects by ignoring therapist effects may lead to results that are not replicable (or valid).

Small effect sizes – overemphasizing small differences

Taking findings from different areas of research into account, Ioannidis (2005b) concluded that the smaller the effect sizes in a scientific field, the less likely the findings are to be true. Small effect sizes, however, may be a replicable result. When comparing, for example, bona fide treatments in psychotherapy research, small differences are rather the rule than the exception (Cuijpers et al. 2013a; Wampold & Imel, 2015). In other cases, however, small differences may just turn out to be sheer randomness or nothing but noise (Ioannidis, 2005b; Wampold & Imel, 2015). Even if they are statistically significant, they may not be clinically relevant. As emphasized by Meehl (1978, p. 822) ‘the null hypotheses, taken literally, is always false’, implying that rejecting the null hypothesis is not a strong test of a substantive hypothesis (Meehl, 1978). The magnitude of the difference is the crucial variable here (Cohen, 1990, p. 1309) ‘because science is inevitably about magnitudes’. Another bias may occur if researchers do not a priori define the difference they are planning to regard as clinically meaningful.
(e.g. \( d \geq 0.25 \)), the post-hoc interpretation of a (small) difference leaves room for arbitrary decisions (e.g. ‘treatment X is superior to Y’), thus constituting a further risk factor of non-replicability. This is especially true if significant but small differences are overemphasized in interpreting research results. A recent meta-analysis on pharmacotherapy and psychotherapy may serve as an example for small effects turning out to be not robust.

Cuijpers and colleagues tested the hypothesis that patients in placebo-controlled trials treated with pharmacotherapy cannot be sure to receive an active drug and may therefore not benefit from the typical and well-documented effects of positive expectancies to the same degree as patients treated with psychotherapy (Cuijpers et al., 2015). The authors hypothesized that (Cuijpers et al., 2015, p. 686) ‘studies that also included a placebo condition (blinded pharmacotherapy) differed significantly from the studies in which no placebo condition was included (unblinded pharmacotherapy)’. When the authors directly compared studies with and without a placebo condition, no significant difference was found for the effects of psychotherapy vs. pharmacotherapy (\( p = 0.15 \)) (Cuijpers et al., 2015, p. 689). Thus, the authors’ hypothesis was not corroborated. The meta-analysis by Cuijpers et al. highlights several problems related to replicability. (a) Despite the insignificant result, Cuijpers et al. performed a secondary analysis comparing the effects of psychotherapy and pharmacotherapy separately for studies with and without a placebo condition. Performing a less strict test when a stricter test (direct comparison) has already failed to corroborate the hypothesis is questionable anyway. For the secondary analysis, the authors reported a non-significant effect (\( g = 0.02 \)) for the first condition (blinded pharmacotherapy) and a significant, but small effect size of \( g = -0.13 \), for the second condition (unblinded pharmacotherapy). They concluded (Cuijpers et al., 2015, p. 691): ‘the results of this study do indicate that blinding in the pharmacotherapy condition reduces the effects’ – which is in contradiction to the first insignificant test reported above. (b) Furthermore, the small effect of \(-0.13\) turned out to be not robust. In a sensitivity analysis by Cuijpers et al. the effects were no longer significant if only CBT was included in the comparison with pharmacotherapy (Cuijpers et al., 2015, p. 690) Thus, the difference of \( g = -0.13 \), which included all forms of psychotherapy, is probably due to the fact that some forms of psychotherapy were less efficacious than CBT (compared to pharmacotherapy), such as non-directive counseling (Cuijpers et al., 2013c). As a consequence, the significant difference found in the authors’ secondary analysis cannot be attributed to unblinding of pharmacotherapy. A more detailed review of this meta-analysis was given elsewhere (Leichsenring et al., 2016).

### Flexibility in design: multiple outcome measures and selective outcome reporting

The more ‘flexibly’ hypotheses and design features are described in the study protocol, the higher the risk for non-replicability (Ioannidis, 2005b). The meta-analysis by Cuijpers et al. (2015) just discussed also highlights the problem of too much flexibility in design, definitions (e.g. of ‘psychotherapy’) and statistical analysis.

The use of multiple outcome measures constitutes a specific problem in that it allows for selective reporting, especially if the primary outcome is not clearly specified. In addition, multiple measures imply problems for statistical testing, particularly type-I error inflation that may lead to overestimating effect sizes (Asendorpf et al., 2016). There is evidence of selective reporting of only favorable results in many areas of research (Chan et al., 2004; Ioannidis, 2005b). As a response to selective reporting, an initiative was established in 2013 called ‘restoring invisible and abandoned trials’ (RIAT, Doshi et al., 2013). Within the RIAT initiative, a study of paroxetine by Keller et al. (2001) on depression in adolescents was recently criticized for selective reporting (Le Noury et al., 2015). The authors reported superiority of paroxetine over placebo; however, this was true only for four outcome measures not pre-specified in the protocol, but not for the primary outcome (Keller et al., 2001, table 2, p. 766).

### Small sample sizes

Small sample size may imply several problems, especially for randomization, generalization, statistical power and, last but not least, for replicability and validity. With regard to randomization, the smaller the study, the less likely pre-existing differences between subjects are randomly distributed between study conditions by randomization (Hsu, 1989), implying a threat to internal validity. In addition, statistical power may be impaired. For instance, among trials comparing psychotherapies for depression, the sample sizes per group in a recent comprehensive meta-analysis ranged between 7 and 113, with a mean sample size per group of 33 (Cuijpers et al., 2013c). Thirty-three subjects per group only allow detection of a relatively large effect size of \( d = 0.70 \) with a power of 0.80 (Cohen, 1988, p. 36). For showing equivalence of a treatment under study to an established treatment with a power of 0.80, a sample size of 33 is not sufficient if smaller margins are accepted as consistent with equivalence (Walker & Nowacki, 2011; Leichsenring et al., 2015b). This result was corroborated by a recent study showing that for psychotherapy of depression more than 100 studies comparing active treatments were recently found to be heavily underpowered (Cuijpers, 2016). As a consequence, if
no significant differences between active treatments are found, equivalence of treatments in outcome may be erroneously concluded (Leichsenring et al. 2015b), a result which may not be replicated by higher powered studies. The relationship between replicability and sample size was recently corroborated by Tajika et al. (2015). The authors reported low rates of replication for studies of pharmacotherapy and psychotherapy, with studies of a total sample size of 100 or more tending to produce replicable results. In psychotherapy research, only a few studies are presently sufficiently powered for demonstrating equivalence or non-inferiority (Leichsenring et al. 2015b; Cuijpers, 2016).

With more than 100 underpowered RCTs only in depression (Cuijpers, 2016), small sample sizes are a common problem.

Meta-analyses can achieve a higher power. In meta-analyses, the statistical power depends on the sample size per study, the number of studies, the heterogeneity between studies, the effect size and the level of significance (Borenstein et al. 2011).

Publication bias

Studies reporting significant effects have a higher likelihood of getting published (Rothstein et al. 2005). However, if non-significant results are not published, the available evidence is distorted. For example, in a meta-analysis of antidepressant medications, Turner et al. (2008) found an effect size of 0.37 for published studies and of 0.15 for unpublished studies. According to two recent meta-analyses, the effects of psychotherapy for depression also seem to be overestimated due to publication bias (Cuijpers et al. 2010; Driessen et al. 2015). Thus, despite being well known, publication bias is still not sufficiently controlled for. Overestimating treatment effects due to publication bias can be expected to reduce both replicability and validity of results. At present, replication or null findings will not receive the same impact as a novel finding and thus will be less helpful to a new scholar’s career progress. So there are disincentives to replication that are built into the whole system. We are in need of a replicability culture.4

Risk factors for non-replicability in meta-analysis

Meta-analyses are based on presently existing studies. Thus, the risk factors for individual studies discussed above necessarily affect the outcome of meta-analyses, too. In addition, the results of meta-analyses heavily depend on the studies that are included or excluded – much as cooking a meal depends on the ingredients you use and the ones you leave out. This fact may have led Eysenck to his provocative ‘garbage-in–garbage-out’ statement about meta-analysis (Eysenck, 1978, p. 517). A recent systematic review corroborated that non-financial conflicts of interest, especially researcher allegiance, are common in systematic reviews of psychotherapy (Lieb et al. 2016). On the other hand, by examining heterogeneity between studies, meta-analyses permit tests of the replicability of results (Asendorpf et al. 2016). Low between-study heterogeneity is indicative of replicability. However, there are a number of ways in which this process of selection may impact the replicability (and validity) of study findings, including the following.

Selectively including studies of non-bona fide treatments in meta-analyses

If studies of non-bona fide treatments are included as comparisons to a specific treatment under investigation, the between-group differences can be expected to be overestimated. This problem may be highlighted by a recent meta-analysis.

Within their meta-analysis on the Dodo bird hypothesis Marcus et al. (2014) compared PDT to CBT. The comparison of PDT to CBT was based on only three included studies of PDT – that is, on a highly selected sample of studies. On the other hand, a large number of bona fide studies were excluded (see the next section). Of these three studies, none can be considered as fully representative of bona fide PDT: In the first study, no treatment manual was used and therapists were not trained for the study (Watzke et al. 2012). In the second study only two plus one sessions were offered to individuals with subsyndromal depression (Barkham et al. 1999). Thus, no sufficient dosage of PDT was applied, and, in addition, no clinical population was treated. Thus, the study by Watzke et al. (2012) and Barkham et al. (1999) do not fulfill the authors’ own inclusion criteria requiring both bona fide treatments and patients (Marcus et al. 2014, p. 522). The third study by Giesen-Bloo et al. (2006) was controversially discussed with regard to the question whether PDT was as carefully implemented as CBT (see above, Giesen-Bloo et al. 2006; Giesen-Bloo & Arntz, 2007; Yeomans, 2007). Thus, in all these three studies, problems with treatment integrity seem to be relevant, yet the conclusions of the meta-analysis were heavily dependent on the findings of these studies.

Selectively excluding studies of bona fide treatments from meta-analyses

If bona fide studies of a treatment are selectively excluded as comparisons to a specific treatment under investigation, between-group differences can be expected to be overestimated. Several meta-analyses may serve as examples.

• The meta-analysis by Marcus et al. (2014) discussed above included only three studies of PDT, but
omitted several RCTs comparing bona fide PDT with other bona fide psychotherapies listed in recent reviews (Leichsenring et al. 2015a, b). Due to this limitation, the meta-analysis by Marcus et al. (2014) cannot claim to be representative of the available evidence for the comparison of bona fide psychotherapies or to provide a valid test of the dodo bird hypothesis.

- Baardseth et al. (2013) noted that several studies of bona fide psychotherapies were excluded in another meta-analysis purporting to find a consistent advantage for a particular family of treatments (Tolin, 2010).

Both including studies using non-bona fide forms of a specific treatment and excluding studies of bona fide treatments can be expected to affect the replicability and validity of meta-analytic results. Meta-analyses that correctly include studies of bona fide treatments can be expected to yield results deviating from those of the above meta-analyses.

Conclusions

The examples reported above suggest that despite considerable efforts several biases are not yet sufficiently controlled for and still affect the quality of published research and its replicability.

There are ‘loopholes’ in the existing standards. For these reasons, we suggest the following measures.

(1) Neutering of treatments may be avoided by specifying, for example, the TIDieR guide (Hoffmann et al. 2014) in a way that deviations of the planned treatment from a clinically established treatment relevant to its efficacy are identified – which is presently not the case.

(2) Researcher allegiance, a powerful risk factor (Luborsky et al. 1999; Falkenström et al. 2013; Munder et al. 2013), has not yet been explicitly addressed in any of the existing guidelines. The CONSORT or PRISMA statements, for example, include items addressing bias of individual studies (Moher et al. 2010, 2015) and meta-biases (such as publication bias) (Moher et al. 2015), but in quite a non-specific way. The respective item of the CONSORT 2010 checklist, for example, states only that researchers should address (Moher et al. 2010, p. 31) ‘trial limitations, addressing sources of potential bias’. It is left to the researcher how to address potential biases. The researchers own allegiance is not mentioned at all. This is also true for the TIDieR guidelines recently developed to improve the replicability of interventions (Hoffmann et al. 2014). The Cochrane Risk of Bias Tool (Higgins et al. 2011) is more explicit in listing several sources of bias (e.g. concealment of allocation, blinding, or selective outcome reporting), but does not address researcher allegiance. For this reason, we make the following suggestions:

- We propose including pertinent items explicitly addressing the researchers own allegiance, for example, in the CONSORT, TIDieR or PRISMA statements or in journal guidelines using indicators established in previous research (Miller et al. 2008; Munder et al. 2012; Lieb et al. 2016). Items such as the following may be helpful: ‘Describe for each treatment condition whether (a) the treatment and/or (b) the associated etiological model was developed and/or (c) advocated by one of the authors, (d) the therapists were trained or supervised by one of the authors, (e) the therapists orientation matches with study condition, (f) the treatments were structurally comparable, for example regarding, duration, dose, or manualization.’ Furthermore, items addressing adversarial collaboration may be added. As illustrated by the examples reported above, the usual statements including the conflict of interest statements are not sufficient here (Lieb et al. 2016).

- Furthermore, researcher bias may be reduced by new methods for data analysis (Miller & Stewart, 2011; MacCoun & Perlmutter, 2015; Nuzzo, 2015; Silberzahn & Uhlmann, 2015), ‘triple-blind’, ‘crowdsourcing’ (see Table 1).

- On an experimental level, researcher allegiance can best be controlled for by including researchers of the different approaches on an equal basis, i.e. an adversarial collaboration (Mellers et al. 2001), both in individual trials and meta-analyses (Nuzzo, 2015). Only by this procedure, design features possibly favoring one’s own approach can really be controlled for. In psychotherapy research, only a few such studies presently exist (e.g. Leichsenring & Leibing, 2003; Gerber et al. 2011; Stangier et al. 2011; Thoma et al. 2012; Leichsenring et al. 2013; Milrod et al. 2015).

(3) Reviewers may be biased in the same way as researchers.

- Reviewer bias may be avoided by new methods for peer review presently discussed, e.g. reviewing a study design prior to knowing the results (Nuzzo, 2015). If the design is approved, the researchers get an ‘in-principle’ guarantee of acceptance, no matter how the results turn out to be (Nuzzo, 2015). Several journals have implemented these procedures (‘registered reports’) or are planning to do so (Nuzzo, 2015).

- Furthermore, some journals (e.g. BMC Psychiatry and other BMC journals) publish the manuscript, and the reviews along with the reviewers’ name on the journal website.
• For grant applications, we are suggesting a comparable procedure to disclose the reviewers’ names, the quality of the reviews and the exact reasons for acceptance/rejection of a proposal.

We hope that our suggestions will contribute to improving replicability in psychotherapy and pharmacotherapy research.

Notes
1 For illustration, we are referring to selected studies that highlight specific risks for replicability. We ask the respective authors to not regard our discussion directed against them or their research. We are aiming at improving the credibility of research. Furthermore, we are aware that we are not free of biases as well.
2 We thank anonymous reviewer no. 1 for making us aware of this problem.
3 We thank anonymous reviewer no. 2 for calling our attention to this issue.
4 We thank anonymous reviewer no. 2 for calling our attention to this issue.
5 Marcus et al. justified the selection of journals by their aim to replicate the 1997 meta-analysis by Wampold et al. (1997). However, in the year 2014 with almost all journal content being available online, there is no need to limit a search for studies to six selected journals.

Declaration of Interest
None.

References
Davidson JR, Foa EB, Huppert JD, Keefe FJ, Franklin ME, Compton JS, Zhao N, Connor KM, Lynch TR, Gadde KM
Fluoxetine, comprehensive cognitive behavioral therapy, and placebo in generalized social phobia. *Archives of General Psychiatry* 61, 1005–1013.


