words: 5389

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

To my late teacher and friend Willi Hager (Falk Leichsenring)

Falk Leichsenring[1], DSc, Allan Abbass[2], MD FRCPC, Mark J. Hilsenroth[3], PhD, Frank Leweke[1], MD, Patrick Luyten[4], PhD, John R. Keefe[5], MA, Nick Midgley[6], PhD, Sven Rabung[7], PhD, Simone Salzer[8], DSc, Christiane Steinert[1], PhD

[1] Department of Psychosomatics and Psychotherapy, Justus-Liebig-University Giessen, Ludwigstr. 76, D-35392 Giessen, Germany

[2] Department of Psychiatry, Dalhousie University; Centre for Emotions and Health, Halifax, 8203 5909 Veterans Memorial Lane, Halifax, NS, Canada, B3H 2E2

[3] The Derner Institute of Advanced Psychological Studies, Adelphi University, Hy Weinberg Center, 1 South Avenue, Garden City, NY 11530-0701, USA

[4] Faculty of Psychology and Educational Sciences, University of Leuven, Klinische Psychologie (OE), Tiensestraat 102 - bus 3722, 3000 Leuven, Belgium, and Research Department of Clinical, Educational and Health Psychology, University College London, Gower Street, London WC1E 6BT, UK

[5] Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA

[6] The Anna Freud Centre, London, and the Research Department of Clinical, Educational and Health Psychology, UCL, Gower Street, London WC1E 6BT, UK

[7] Department of Psychology, Alpen-Adria-Universität Klagenfurt, Universitätsstr. 65-67, A-9020 Klagenfurt, Austria, and Department of Medical Psychology, University Medical Center Hamburg-Eppendorf, Martinistr. 52, D-20246 Hamburg, Germany

[8] Clinic of Psychosomatic Medicine and Psychotherapy, Georg-August-Universität Goettingen Von-Sieboldstr. 5, D-37075 Göttingen, Germany and International Psychoanalytic University (IPU) Berlin, Germany

Abstract

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 thirteen biases as a starting point, we discuss each bias´ 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 of findings. 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).

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

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 (Asendorpf *et al.*, 2016, Brunswik, 1955). 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 to 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, Ioannidis *et al.*, 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 nonreplicability is important. Biases in research are well-known for affecting results (e.g. Ioannidis, 2005b). 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.

Methods

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, 2005b, p. 0697). We used a list of well-known biases made up by Ioannidis (2005b) 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 and 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.[1] We did not aim at a examining a random sample of studies, but rather chose to highlight the relevance of these risk factors by demonstrative examples.

- insert Table 1 about here -

Results

1. Allegiance effects

1.1 Researcher allegiance

In biomedical research, conflicts of interest and prejudice are common, but only sparsely reported, let alone controlled for (Dragioti *et al.*, 2015, Ioannidis, 2005b). In psychotherapy

[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.

research, researcher's own allegiances 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 (Falkenström *et al.*, 2013, Munder *et al.*, 2012), 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 and Imel, 2015). This was the case, for example in the 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 hours of role-playing. (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 1.2,), which may be additionally fostered by the differences in training duration. (3) 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 6 of these 39 tests, a statistically significant difference in outcome in favor of CT was found. It is not known how many of them are due to chance. (4) 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.

1.2 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 and Imel, 2015, p. 170). These results suggest that therapist allegiance may play an important role in pharmacotherapy as well.

1.3. 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 researcher, therapist and/or supervisor allegiance were discussed, for example, by Wampold and Imel (2015 , p. 120-128). Measures to control for allegiance effects are proposed below (Conclusions and Table 1).

1.4. 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, 2005b).

1.5. 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).

2. Impaired treatment integrity: "strawman" therapies

Treatment integrity is defined as the degree to which treatments are carried out as originally intended (Kazdin, 1994, Yeaton and Sechrest, 1981). – 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. Davidson *et al.*, 2004, p. 1006, Elkin, 1985), 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 and 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 nonbona-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.[2]

[2] We thank the anonymous reviewer # 1 for making us aware of this problem.

An RCT comparing PDT to CBT in adolescents with 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 two days, whereas the CBT therapists were trained for five 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 important, 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 and 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 nonspecific treatment components. In an RCT by Snyder and 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, nonspecific 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 and 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.

3. 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 and 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 and Imel, 2015). For this reason,

therapists need to be statistically taken into account as a nested random factor (Wampold and Imel, 2015), although larger sample sizes are needed to achieve this (Wampold and Imel, 2015). Failure do so may result in increased type I errors and overestimating treatment effects (Wampold and 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 and 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).

4. 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 and Imel, 2015). In other cases, however, small differences

may just turn out to be sheer randomness or nothing but noise (Ioannidis, 2005b, Wampold and 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 counselling.(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).

5. 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-analyses 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).

6. 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.*, 2013b). 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 (Leichsenring *et al.*, 2015b, Walker and Nowacki, 2011). 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 (Cuijpers, 2016, Leichsenring *et al.*, 2015b).

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 and the effect size and the level of significance (Borenstein *et al.*, 2011).

7. 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.[3] We are in need of a replicability culture.[4]

8. Risk factors for nonreplicability 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.

8.1 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

[3] We thank the anonymous reviewer # 2 for calling our attention to this issue.

[4] We thank the anonymous reviewer # 2 for calling our attention to this issue.

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 below, 8.2). 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 studies by Watzke *et al.* (2012) and Barkham *et al.* (1999) do not fulfil 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 and Arntz, 2007, Giesen-Bloo *et al.*, 2006, 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.

8.2 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, Leichsenring *et al.*, 2015b).[5] 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.

[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.

(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 (Falkenström *et al.*, 2013, Luborsky *et al.*, 1999, 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, Moher *et al.*, 2015) and meta-biases (such as publication bias)(Moher *et al.*, 2015), but in a 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 (Lieb *et al.*, 2016, Miller *et al.*, 2008, Munder *et al.*, 2012). 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 (MacCoun and Perlmutter, 2015, Miller and Stewart, 2011, Nuzzo, 2015, Silberzahn and Uhlmann, 2015), “triple-blind”, “croudsourcing”, see Table1).
- 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. Gerber *et al.*, 2011, Leichsenring and Leibing, 2003, Leichsenring *et al.*, 2013, Milrod *et al.*, 2015, Stangier *et al.*, 2011, Thoma *et al.*, 2012).

(2) 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 researches get an “in-principle” guarantee of acceptance, no matter how the results turn out to be (Nuzzo, 2015). Several journals have implemented these procedures or are planning to do so (Nuzzo, 2015).
- Furthermore, some journals (e.g. BMC Psychiatry and other BioMed Central 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 revewers´ 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.

Table 1: Proposed measures to control for fisk factors for non-replicability in psychotherapy and pharmacotherapy

	Risk Factors	**Proposed Measures**
1.	Allegiances	
1.1	Researcher Allegiance	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.
1.2	Therapist allegiance	Treatments are carried out by experts of the respective approach, therapists do not carry out treatments A, B or C being compared. The same applies to treatment supervisors.
1.3	Supervisor allegiance	Therapists are supervised by experts in the respective approach. No supervision of different treatments by the same therapist.
1.4	Reviewer Allegiance	Blinded reviewers; review of study design prior to knowing the results; no anonymous reviews; public control of reviewer decisions (especially for grant applications).
1.5	Editor Allegiance/Policy	Editor allegiance may be reduced by measures for a more open and transparent journal policy, e.g. registered reports.
2	Impaired treatment integrity ("strawman" therapies)	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).
3	Ignoring therapist effects	Taking therapist effects in data analysis systematically into account; report of effect sizes for therapist effects (ICC).
4	Small effect sizes: overemphasizing small differences	Differentiating between statistically and clinically significant findings; a priori defining a clinically meaningful threshold in upfront trial registration.
5	Flexibility in design: multiple outcome measures and selective outcome reporting	Upfront study registration including primary and secondary outcomes; focus on ITT analyses.
6	Small sample sizes	Performing higher powered studies when addressing relatively established findings; meta-analyses achieve higher power than small individual studies.
7	Publication Bias	Upfront trial registration; increased publication of non-significant results (change in editor policy), acceptance of manuscripts before results are known.

8.1	Selective inclusion of non-bona fide studies in meta-analyses	Upfront registration; measures described above to control for allegiances of researchers, reviewers and editors.
8.2	Selective exclusion of bona fide studies in meta-analyses	Upfront registration, measures described above to control for allegiances of researchers, reviewers and editors.

References

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F. & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *Journal of the American Medical Association* **268**, 240-8.

Asendorpf, J., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J., Fiedler, K., Fiedler, S., Funder, D. C., Kliegel, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H. & Wicherts, J. M. (2016). Recommendations for increasing replicability in psycholgy. In *Methodological issues and strategies in clinical research (4th ed.)* (ed. A. Kazdin), pp. 607-622. American Psychological Association Washington, DC, US

Baardseth, T. P., Goldberg, S. B., Pace, B. T., Wislocki, A. P., Frost, N. D., Siddiqui, J. R., Lindemann, A. M., Kivlighan, D. M., 3rd, Laska, K. M., Del Re, A. C., Minami, T. & Wampold, B. E. (2013). Cognitive-behavioral therapy versus other therapies: redux. *Clinical Psychology Review* **33**, 395-405.

Barkham, M., Shapiro, D. A., Hardy, G. E. & Rees, A. (1999). Psychotherapy in two-plus-one sessions: Outcomes of a randomized controlled trial of cognitive-behavioral and psychodynamic-interpersonal therapy for subsyndromal depression *Journal of Consulting and Clinical Psychology* **67** 201-211.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2011). *Introduction to meta-analysis*. Wiley: Chichester, UK.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review* **62**, 193-217.

Carey, B. (2015). Psychology´s fears confirmed: Rechecked studies don´t hold up. *New York Times 27. August 2015*, A1.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., Gotzsche, P. C. & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* **291**, 2457-65.

Clark, D. M., Salkovskis, P. M., Hackmann, A., Middleton, H., Anastasiades, P. & Gelder, M. (1994). A comparison of cognitive therapy, applied relaxation and imipramine in the treatment of panic disorder. *British Journal of Psychiatry* **164**, 759-69.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist* **45**, 1304-1312.

Cottraux, J., Note, I. D., Boutitie, F., Milliery, M., Genouihlac, V., Yao, S. N., Note, B., Mollard, E., Bonasse, F., Gaillard, S., Djamoussian, D., Guillard Cde, M., Culem, A. & Gueyffier, F. (2009). Cognitive therapy versus Rogerian supportive therapy in borderline personality disorder. Two-year follow-up of a controlled pilot study. *Psychotherapy and Psychosomatics* **78**, 307-16.

Cuijpers, P. (2016). Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evidence-Based Mental Health*.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A. & Dobson, K. S. (2013a). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry* **58**, 376-85.

Cuijpers, P., Huibers, M., Ebert, D. D., Koole, S. L. & Andersson, G. (2013b). How much psychotherapy is needed to treat depression? A metaregression analysis. *Journal of Affective*

Disorders **149**, 1-13.
Cuijpers, P., Karyotaki, E., Andersson, G., Li, J., Mergl, R. & Hegerl, U. (2015). The effects of blinding on the outcomes of psychotherapy and pharmacotherapy for adult depression: A meta-analysis. *European Psychiatry* **30**, 685-93.
Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T. & Reynolds, C. F., 3rd (2013c). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons. *World Psychiatry* **12**, 137-48.
Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D. & Andersson, G. (2010). Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *British Journal of Psychiatry* **196**, 173-8.
Davidson, J. R., Foa, E. B., Huppert, J. D., Keefe, F. J., Franklin, M. E., Compton, J. S., Zhao, N., Connor, K. M., Lynch, T. R. & Gadde, K. M. (2004). Fluoxetine, comprehensive cognitive behavioral therapy, and placebo in generalized social phobia. *Archives of General Psychiatry* **61**, 1005-13.
Doshi, P., Dickersin, K., Healy, D., Vedula, S. S. & Jefferson, T. (2013). Restoring invisible and abandoned trials: a call for people to publish the findings. *BMJ (Clinical Research Ed.)* **346**, f2865.
Dragioti, E., Dimoliatis, I. & Evangelou, E. (2015). Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. *BMJ Open* **5**, e007206.
Driessen, E., Hollon, S. D., Bockting, C. L., Cuijpers, P. & Turner, E. H. (2015). Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-Funded Trials. *PLoS ONE* **10**, e0137864.
Dush, D. M., Hirt, M. L. & Schroeder, H. (1983). Self-statement modification with adults: a meta-analysis. *Psychological Bulletin* **94**, 408-422.
Elkin, I., Parloff, MB, Hadley, SW, Autry, J.H. (1985). NIMH Treatment of Depression collaborative Research Program. *Archives of General Psychiatry* **42**, 305-316.
Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P. & et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Archives of General Psychiatry* **46**, 971-983.
Eysenck, H. J. (1978). An exercise in mega-silliness *American Psychologist* **33**, 517.
Falkenström, F., Markowitz, J. C., Jonker, H., Philips, B. & Holmqvist, R. (2013). Can psychotherapists function as their own controls? Meta-analysis of the crossed therapist design in comparative psychotherapy trials. *Journal of Clinical Psychiatry* **74**, 482-91.
Foa, E. B., Rothbaum, B. O., Riggs, D. S. & Murdock, T. B. (1991). Treatment of posttraumatic stress disorder in rape victims: a comparison between cognitive-behavioral procedures and counseling. *Journal of Consulting and Clinical Psychology* **59**, 715-23.
Fugelsang, J. A., Stein, C. B., Green, A. E. & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology* **58**, 86-95.
Gerber, A. J., Kocsis, J. H., Milrod, B. L., Roose, S. P., Barber, J. P., Thase, M. E., Perkins, P. & Leon, A. C. (2011). A quality-based review of randomized controlled trials of psychodynamic psychotherapy. *American Journal of Psychiatry* **168**, 19-28.
Giesen-Bloo, J. & Arntz, A. (2007). Questions concerning the randomized trial of schema-focused therapy vs transference-focused psychotherapy – Reply *Archives of General Psychiatry* **64**, 610-611.

Giesen-Bloo, J., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., van Asselt, T., Kremers, I., Nadort, M. & 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.

Gilboa-Schechtman, E., Foa, E. B., Shafran, N., Aderka, I. M., Powers, M. B., Rachamim, L., Rosenbach, L., Yadin, E. & Apter, A. (2010). Prolonged exposure versus dynamic therapy for adolescent PTSD: a pilot randomized controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry* **49**, 1034-42.

Higgins, J., P.,, Altman, D. G., Goetzsche, P. C., Juni, P., Moher, D. & Oxman, A. D. (2011). The Cochrane Statistical Methods Group. The Cochrane Collaboration´s tool for assesingrisk of bias in randomized trials. *British Medical Journal* **343**, d5928.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A. W. & Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Clinical Research Ed.)* **348**, g1687.

Horowitz, M. & Kaltreider, N. (1979). Brief therapy of the stress response syndrome. *Psychiatric Clinics of North America* **2**, 365-377.

Hsu, L. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research *Journal of Consulting and Clinical Psychology* **57**, 131-137.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *JAMA* **294**, 218-28.

Ioannidis, J. P. (2005b). Why most published research findings are false. *PLoS Medicine* **2**, e124.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E. & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics* **41**, 149-55.

Jacobson, N. S. (1991). To be or not to be behavioral when working with couples. What does it mean ? *Journal of Familiy Psychology* **4**, 436-445.

Kazdin, A. E. (1994). Methodology, design, and evaluation in psychotherapy research. In *Handbook of psychotherapy and behavior change* (ed. A. E. Bergin and S. L. Garfield), pp. 19-71. Wiley: New York.

Keller, M. B., Ryan, N. D., Strober, M., Klein, R. G., Kutcher, S. P., Birmaher, B., Hagino, O. R., Koplewicz, H., Carlson, G. A., Clarke, G. N., Emslie, G. J., Feinberg, D., Geller, B., Kusumakar, V., Papatheodorou, G., Sack, W. H., Sweeney, M., Wagner, K. D., Weller, E. B., Winters, N. C., Oakes, R. & McCafferty, J. P. (2001). Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry* **40**, 762-72.

Le Noury, J., Nardo, J. M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C. & Abi-Jaoude, E. (2015). Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ (Clinical Research Ed.)* **351**, h4320.

Leichsenring, F. & Leibing, E. (2003). The effectiveness of psychodynamic therapy and cognitive behavior therapy in the treatment of personality disorders: a meta-analysis. *American Journal of Psychiatry* **160**, 1223-32.

Leichsenring, F., Leweke, F., Klein, S. & Steinert, C. (2015a). The empirical status of psychodynamic psychotherapy - an update: Bambi's alive and kicking. *Psychotherapy and Psychosomatics* **84**, 129-48.

Leichsenring, F., Luyten, P., Hilsenroth, M. J., Abbass, A., Barber, J. P., Keefe, J. R., Leweke, F., Rabung, S. & Steinert, C. (2015b). Psychodynamic therapy meets evidence-based medicine: a systematic review using updated criteria. *Lancet Psychiatry* **2**, 648-60.
Leichsenring, F., Salzer, S., Beutel, M. E., Herpertz, S., Hiller, W., Hoyer, J., Huesing, J., Joraschky, P., Nolting, B., Poehlmann, K., Ritter, V., Stangier, U., Strauss, B., Stuhldreher, N., Tefikow, S., Teismann, T., Willutzki, U., Wiltink, J. & Leibing, E. (2013). Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: a multicenter randomized controlled trial. *American Journal of Psychiatry* **170**, 759-67.
Leichsenring, F., Steinert, C. & Hoyer, J. (2016). Psychotherapy Versus Pharmacotherapy of Depression: What's the Evidence? *Zeitschrift fur Psychosomatische Medizin und Psychotherapie* **62**, 190-5.
Lieb, K., Osten-Sacken, J., Stoffers-Winterling, J., Reiss, N. & Barth, J. (2016). Conflicts of interest and spin in reviews of psychological therapies: a systematic review. *BMJ Open* **6**, e010606.
Luborsky, L., Diguer, L., Seligman, D., Rosenthal, R., Krause, E., Johnson, S., Halperin, G, , Bishop, M., Berman, J. & Schweizer, E. (1999). The researcher's own allegiances: A 'wild' card in comparison of treatment efficacy *Clinical Psychology: Science and Practice* **6** 95-106.
MacCoun, R. & Perlmutter, S. (2015). Hide resuls to seek the truth. *Nature* **526**, 187-189.
Marcus, D. K., O'Connell, D., Norris, A. L. & Sawaqdeh, A. (2014). Is the Dodo bird endangered in the 21st century? A meta-analysis of treatment comparison studies. *Clinical Psychology Review* **34**, 519-530.
McKay, K. M., Imel, Z. E. & Wampold, B. E. (2006). Psychiatrist effects in the psychopharmacological treatment of depression. *Journal of Affective Disorders* **92**, 287-90.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psycholgy. *Journal of Consulting and Clinical Psychology* **46**, 806-834.
Mellers, B., Hertwig, R. & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science* **12**, 269-75.
Miller, L. E. & Stewart, M. E. (2011). The blind leading the blind: use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials* **32**, 240-3.
Miller, S., Wampold, B. & Varhely, K. (2008). Direct comparisons of treatment modalities for youth disorders: a meta-analysis. *Psychotherapy Research* **18**, 5-14.
Milrod, B., Chambless, D. L., Gallop, R., Busch, F. N., Schwalberg, M., McCarthy, K. S., Gross, C., Sharpless, B. A., Leon, A. C. & Barber, J. P. (2015). Psychotherapies for panic disorder: a tale of two sites. *Journal of Clinical Psychiatry*.
Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M. & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal (Clinical Research Ed.)* **340**, c869.
Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews* **4**, 1.
Munder, T., Brutsch, O., Leonhart, R., Gerger, H. & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical Psychology Review* **33**, 501-11.
Munder, T., Fluckiger, 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**, 631-7.
Munder, T., Gerger, H., Trelle, S. & Barth, J. (2011). Testing the allegiance bias hypothesis: a

meta-analysis. *Psychotherapy Research* **21**, 670-84.
Nuzzo, R. (2015). How scientists fool themselves - and how they can stop. *Nature* **526**, 182-5.
Open Science Collaboration (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716.
Owen, J., Drinane, J. M., Idigo, K. C. & Valentine, J. C. (2015). Psychotherapist effects in meta-analyses: How accurate are treatment effects? *Psychotherapy (Chic)* **52**, 321-8.
Perepletchikova, F., Treat, A. T. & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology* **75**, 829-841.
Popper, K. R. (1959). *The logic of scientific discovery*. Basic Books, 2014: New York.
Rothstein, H. R., Sutton, A. J. & Borenstein, M. (2005). Publication bias. In *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustment* (ed. H. R. Rothstein, A. J. Sutton and M. Borenstein), pp. 277-302. Wiley & Sons: New York.
Silberzahn, R., . & Uhlmann, E. L. (2015). Many hands make tight work. *Nature* **526**, 189-191.
Smith, M. L., Glass, G. V. & Miller, T. I. (1980). *The benefits of psychotherapy*. John Hopkins University Press: Baltimore.
Snyder, D. K. & Wills, R. M. (1989). Behavioral vs. insight-oriented marital therapy: Effects on individual and interspousal functioning. *Journal of Consulting and Clinical Psychology* **57**, 39-46.
Snyder, D. K., Wills, R. M. & Grady-Fletcher, A. (1991). Long-term effectiveness of behavioral versus insight-oriented marital therapy: a 4-year follow-up study. *Journal of Consulting and Clinical Psychology* **59**, 138-41.
Spiwak, M. (2016). Nothing but reviews in mind [Nichts als Gutachten im Kopf]. *DIE ZEIT* **32**, 31-32.
Stangier, U., Schramm, E., Heidenreich, T., Berger, M. & Clark, D. M. (2011). Cognitive therapy vs interpersonal psychotherapy in social anxiety disorder: a randomized controlled trial. *Archives of General Psychiatry* **68**, 692-700.
Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y. & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *British Journal of Psychiatry* **207**, 357-62.
Thoma, N. C., McKay, D., Gerber, A. J., Milrod, B. L., Edwards, A. R. & Kocsis, J. H. (2012). A quality-based review of randomized controlled trials of cognitive-behavioral therapy for depression: an assessment and metaregression. *American Journal of Psychiatry* **169**.
Tolin, D. F. (2010). Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review* **30**, 710-20.
Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* **358**, 252-60.
Walker, E. & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine* **26**, 192-6.
Wampold, B. E. & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. Routledge: New York.
Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K. & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes". *Psychological Bulletin* **122**, 203-215.
Watzke, B., Rüddel, H., Jürgensen, R., Koch, U., Kriston, L., Grothgar, B. & Schulz, H. (2012). Longer term outcome of cognitive-behavioural and psychodynamic psychotherapy in routine mental health care: Randomised controlled trial. *Behaviour Research and Therapy* **50**

580-587.
Woeller, W., Leichsenring, F., Leweke, F. & Kruse, J. (2012). Psychodynamic psychotherapy for posttraumatic stress disorder related to childhood abuse. - Principles for a treatment manual. *Bulletin of the Menninger Clinic* **76**, 69-93.
Yeaton, W. H. & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology* **49**, 156-167.
Yeomans, F. (2007). Questions concerning the randomized trial of schema-focused therapy vs transference-focused psychotherapy. *Archives of General Psychiatry* **64**, 610-611.