Equivalence testing: reversed hypotheses, margins, and the need for controlling researcher allegiance

Falk Leichsenring[1], DSc, Allan Abbass[2], MD FRCPC, Mark Hilsenroth[3], PhD, Patrick Luyten[4], PhD, Sven Rabung[5], PhD, Christiane Steinert[1,6], PhD

[1] Department of Psychosomatics and Psychotherapy, Justus-Liebig-University Giessen, Ludwigstr. 76, 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] Derner School of Psychology, 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, Alpen-Adria-Universität Klagenfurt, Universitätsstr. 65-67, A-9020 Klagenfurt, Austria

[6] MSB Medical School Berlin, Department of Psychology, Calandrellistr. 1-9, 12247 Berlin, Germany

In press: Psychological Medicine

Correspondig Author:
Prof. Dr. Falk Leichsenring
University of Giessen
Department of Psychosomatics and Psychotherapy
Ludwigstr. 76, 35392 Giessen, Germany
Fon: +49-641-99 45660
Fax: +49-641-99 45664
Mail: Falk.Leichsenring@psycho.med.uni-giessen.de

The number of non-inferiority or equivalence trials is increasing in many areas of research (Piaggio *et al.*, 2012). Thus, the specific statistical and methodological issues involved in these type of trials become more and more important. A recent correspondence article by Rief and Hofmann (2018a) suggests that several of these issues need some clarification.

In a reply to our comment on their article on non-inferiority testing (Leichsenring *et al.*, 2018a, Rief and Hofmann, 2018b), Rief and Hofmann (2018a) reject our statement that they misinterpreted the results of the Steinert et al. meta-analysis (Steinert *et al.*, 2017). They maintain that a significant disadvantage of psychodynamic therapy compared to other therapies was shown. However, statistical results cannot be detached from the scientific hypothesis tested, that is from their context conditions. Steinert et al. tested the hypothesis of equivalence of psychodynamic therapy to other treatments established in efficacy including CBT. In equivalence testing, the null and alternative hypothesis are reversed and a significant results indicates that the null hypothesis of non-equivalence (e.g. $\delta > 0.25$) is rejected and equivalence can be concluded Walker and Nowacki (2011). Thus, the p values (e.g. p=0.016) reported by Steinert *et al.* (2017) do not indicate a disadvantage of psychodynamic therapy as erroneously stated by Rief and Hofmann (2018a, p. 2) but equivalence. Equivalence was convincingly demonstrated since all confidence intervals were completely included within the pre-defined margin (δ) of equivalence ($g= \pm 0.25$), one of the smallest margins ever proposed for demonstrating equivalence in psychotherapy research (Steinert *et al.*, 2017).

Rief and Hofmann do not suggest a realistic and scientifically justified margin for testing equivalence, neither in their first paper (Rief and Hofmann, 2018b) nor in their new article (Rief and Hofmann, 2018a) In their recent paper (Rief and Hofmann, 2018a, p. 2), they just claim that a therapy that is "10% or 20% less effective than the current gold standard”

cannot be recommended as a treatment, neither ethically nor scientifically. This statement is questionable for several reasons addressed in the following.

Rief and Hofmann (2018a) do not give any scientific justification for their proposal: The authors do not specify what they regard as compatible with equivalence, but state only implicitly what they regard as not compatible with equivalence, a difference of 10% or above. Why 10%, not 15% or 5% ?

As an aside: Steinert *et al.* (2017) found a difference between psychodynamic therapy and CBT in terms of Hedges´ g of 0.16 which corresponds to a difference in success rates of less than 9% (Kraemer and Kupfer, 2006). Thus, it is below 10% which is apparently regarded by Rief and Hofmann as compatible with equivalence.

Furthermore, what would a statistically significant difference suggested by Rief and Hofmann (2018a) imply ? Cohen (1988) proposed to consider an effect size of 0.20 as small. Thus, 0.16 is clearly a small effect. In a meta-analysis even very small effect sizes may become statistically significant due to its high statistical power. Such small differences may not be of any clinical significance (McGlothlin and Lewis, 2014, Turner *et al.*, 2010). Rief and Hofmann (2018a) suggest that CBT is superior to psychodynamic therapy by an effect size of 0.16. This corresponds to a difference of about 1 scale point on the Hamilton Depression Rating Scale, for instance, a difference that can hardly be considered to be clinically important. Not only statistical significance, but also clinical significance needs to be into account when defining margins, reviewing results or giving treatment recommendations.

Furthermore, it is not clear what the ‘gold standard’ mentioned by Rief and Hofmann (2018a, p. 2) is, as response rates of most treatments for common mental disorders hover

around 50% with no significant differences between types of psychotherapy (Cuijpers *et al.*, 2014). As we have argued before, there is much room for improvement of current treatments and we are far from having established a 'gold standard' treatment. Specifically the evidence base for CBT is less solid than previously thought if response and remission rates, comparative efficacy, evidence on mechanisms of change and study quality are taken into account (Leichsenring *et al.*, 2018b, Leichsenring and Steinert, 2017).

If the advantage of CBT is not more than about 1 scale point on the Hamilton Depression Rating Scale, for instance, or 0.16 in general despite many years of research and millions of dollars, pounds and euros used to fund research of CBT (MQ, 2015), this is not very impressive and again hardly qualifies as a gold standard. At follow-up the difference between psychodynamic therapy and CBT was still even smaller (g=- 0.05, p=0.0001).

CBT research has been much more funded than research in other approaches: in the UK alone, CBT research was awarded 30.42 million pounds between 2008 and 2013 (MQ, 2015). In the same time interval, psychodynamic therapy was funded with only 1.53 million pounds (MQ, 2015), one twentieth that of CBT. Thus, "equivalence" in funding does not exist. If the starting conditions of treatments are that different, fair comparisons are hardly possible, neither in equivalence nor in superiority trials. For this reason, differences in funding should be paid more attention. As bias in funding decisions were found to be common, they need to be better controlled for (Nicholson and Ioannidis, 2012).

Furthermore, Rief and Hofmann (2018a) question the results of a recent equivalence meta-analysis (Steinert *et al.*, 2017) by repeating their allegation (Rief and Hofmann, 2018a, b) that this meta-analysis was biased as it was supported by a German psychodynamic society. However, in order to control for researcher allegiance, Steinert et al. included a prominent

CBT researcher (JH) in this meta-analysis of psychodynamic therapy, thus establishing a form of adversarial collaboration. As stated in the original article and in a first reply to Rief and Hofmann, the sponsor did not influence the procedures, the results or the presentation of the Steinert at al. meta-analysis in any way (Leichsenring *et al.*, 2018a, Steinert *et al.*, 2017). Thus, it is not clear why Rief and Hofmann keep on repeating their false allegation. Until now Rief and Hofmann have not yet include a prominent psychodynamic researcher in any of their studies to balance any implicit or explicit researcher biases. We invite them to do so.

As research in psychodynamic therapy is rarely funded by public funding organizations not only in the UK but also by the NIMH in the US or the DFG in Germany, psychodynamic researchers are in need for other sponsors - which is then turned against them by CBT advocates, for example, by Rief and Hofmann (2018a) – a vicious circle.

Finally, Rief and Hofmann (2018a, p. 2) claim that a therapy that is "10% or 20% less effective than the current gold standard" cannot be recommended as a treatment, neither ethically nor scientifically. However, this claim would only be justified if there was another therapy with (close to) 100% efficacy. Response rates for the different forms of psychotherapy including CBT, which seems to be regarded as the gold standard by Rief and Hofmann (2018a), are about 50% and rates for remission are even lower (Cuijpers *et al.*, 2014, Loerinc *et al.*, 2015). Thus, many patients do not sufficiently benefit from one specific psychotherapeutic approach. They may, however, benefit from another approach, e.g. from interpersonal therapy, systemic therapy or psychodynamic therapy. The chance of being helpful for the other half of patients should outweigh a small difference in effect sizes such as 0.16. For this reason, recommending only one evidence-based approach - due to a possible advantage of g=0.16 - and excluding others is not justified, neither ethically nor scientifically.

Treatment recommendations need to take the whole complexity of evidence into account instead of focusing on one isolated statistical parameter only.

These considerations suggest the following conclusions for equivalence testing in particular and for research in general.

- Equivalence margins need to be explicitly formulated and scientifically justified, taking the conditions of the respective disorder into account.
- Not only statistical significance, but also clinical significance needs to be into account when defining margins, reviewing results or giving treatment recommendations.
- Researcher allegiance needs to be taken more explicitly into account, in equivalence trials, but also in superiority trials and meta-analyses, as well as in treatment guideline committees, for example by establishing forms of adversarial collaboration.
- Bias´ in funding decisions are common and need to be better controlled for, too.
- For fair comparisons between treatments in equivalence and superiority trials the different treatments need to be more equally supported by funding organizations,
- For the sake of our patients funding in psychotherapy needs to be distributed more broadly. Patients who do not benefit from one evidence-based approach may benefit from another. A diversity of equally-funded and evidence-based treatments is a strength, a psychotherapy monoculture is not.
- Treatment recommendations need to be balanced, taking the whole complexity of evidence into account. Focusing on one isolated statistical parameter only is not justified, neither ethically nor scientifically.

References

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

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D. & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *J Affect Disord* **159**, 118-26.

Kraemer, H. C. & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry* **59**, 990-6

Leichsenring, F., Abbass, A., Driessen, E., Hilsenroth, M., Luyten, P., Rabung, S. & Steinert, C. (2018a). Equivalence and non-inferiority testing in psychotherapy research. *Psychological Medicine* **48**, 1917-1919.

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Luyten, P., Munder, T., Rabung, S. & Steinert, C. (2018b). "Gold Standards," Plurality and Monocultures: The Need for Diversity in Psychotherapy. *Front Psychiatry* **9**, 159.

Leichsenring, F. & Steinert, C. (2017). Is Cognitive Behavioral Therapy the Gold Standard for Psychotherapy?: The Need for Plurality in Treatment and Research. *JAMA* **318**, 1323-1324.

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J. & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review* **42**, 72-82.

McGlothlin, A. E. & Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *JAMA* **312**, 1342-3.

MQ (2015). MQ Landscape Analysis. UK Mental Health Research Funding.

Nicholson, J. M. & Ioannidis, J. P. (2012). Research grants: Conform and be funded. *Nature* **492**, 34-6.

Piaggio, G., Elbourne, D. R., Pocock, S. J., Evans, S. J. & Altman, D. G. (2012). Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA* **308**, 2594-604.

Rief, W. & Hofmann, S. G. (2018a). The limitations of equivalence and non-inferiority trials. *Psychological Medicine*, 1-2.

Rief, W. & Hofmann, S. G. (2018b). Some problems with non-inferiority tests in psychotherapy research: psychodynamic therapies as an example. *Psychological Medicine*, 1-3.

Steinert, C., Munder, T., Rabung, S., Hoyer, J. & Leichsenring, F. (2017). Psychodynamic Therapy: As Efficacious as Other Empirically Supported Treatments? A Meta-Analysis Testing Equivalence of Outcomes. *American Journal of Psychiatry* **174**, 943-953.

Turner, D., Schunemann, H. J., Griffith, L. E., Beaton, D. E., Griffiths, A. M., Critch, J. N. & Guyatt, G. H. (2010). The minimal detectable change cannot reliably replace the minimal important difference. *Journal of Clinical Epidemiology* **63**, 28-36.

Walker, E. & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *J Gen Intern Med* **26**, 192-6.