
Article:
Schiefele, A.K., Lutz, W., Barkham, M. et al. (8 more authors) (2017) Reliability of Therapist Effects in Practice-Based Psychotherapy Research: A Guide for the Planning of Future Studies. Administration and Policy in Mental Health and Mental Health Services Research, 44 (5). pp. 598-613. ISSN 0894-587X

https://doi.org/10.1007/s10488-016-0736-3

Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies

Anne-Katharina Schiefele

University of Trier

Wolfgang Lutz

University of Trier

Michael Barkham

University of Sheffield

Julian Rubel

University of Trier

Jan Böhnke

Mental Health and Addiction Research Group, Hull York Medical School & Department of Health Sciences, University of York

Jaime Delgadillo

Leeds Community Healthcare NHS Trust and Department of Health Sciences, University of York

Mark Kopta

University of Evansville

Dietmar Schulte

Ruhr-Universität Bochum

David Saxon

University of Sheffield

Stevan L. Nielsen

Brigham Young University

Michael J. Lambert

Brigham Young University

Please be aware that this is a pre-publication manuscript!

[http://link.springer.com/article/10.1007%2Fs10488-016-0736-3]

Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., ... & Lambert, M. J. (2016). Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies. *Administration and Policy in Mental Health and Mental Health Services Research*. Advance online publication. doi: 10.1007/s10488-016-0736-3

Corresponding author:

M. Sc. Anne-Katharina Schiefele

Clinical Psychology and Psychotherapy, Department of Psychology, University of Trier

D-54286 Trier, Germany

Phone: +49-651-201-2882; Fax: +49-651-201-2886; E-mail: schiefele@uni-trier

Abstract

This paper aims to provide researchers with practical information on sample sizes for accurate estimations of therapist effects (TEs). The investigations are based on an integrated sample of 48,648 patients treated by 1,800 therapists. Multilevel modeling and resampling were used to realize varying sample size conditions to generate empirical estimates of TEs. Sample size tables, including varying sample size conditions, were constructed and study examples given. This study gives an insight into the potential size of the TE and provides researchers with a practical guide to aid the planning of future studies in this field.

Keywords: Therapist effects, naturalistic data, multilevel analysis, sample size, practical guide

Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies

Although the central role of therapists within the process of psychotherapy is obvious, the contribution of the individual therapist to the variability in treatment outcomes has often been neglected in study designs and analysis (Baldwin & Imel, 2013; Beutler et al., 2004; Garfield, 1997; Lutz & Barkham, 2015). Ricks (1974) reported the first empirical evidence for existing differences between therapists in his "Supershrink" study and the body of literature attesting to differences between therapists has steadily grown (for a review, see Baldwin & Imel, 2013). Based on a narrative review, Lambert (1992) attempted to attribute outcome in psychotherapy to various factors including the patient, the type of therapy and the specific therapist. The results emphasized the importance of the therapist variable to patient outcome and stimulated further investigations.

Crits-Christoph and colleagues (1991) reported the first meta-analysis of therapist effects (TEs) and reanalyzed data from 15 studies and 27 treatment groups extracting an overall TE of 8.6% (Crits-Christoph et al., 1991). Twenty years later, in their review of TEs, Baldwin and Imel (2013) conducted a meta-analysis with more than twice as many studies (n = 46) that showed approximately 5% of the variance in outcomes to be attributable to therapists. However, the percentage differed as a function of research design with only about 3% of the variance associated with the person who delivered the treatment occurring in randomized controlled efficacy studies, but 7% in naturalistic study designs. The utilization of manuals appears to reduce the variance associated with therapists, but there is a debate as to how much reduction in the size of TEs can be explained by the standardization of treatments utilizing manuals (Baldwin & Imel, 2013; Crits-Christoph et al., 1991; Hofmann & Barlow, 2014).

The extant literature would therefore indicate that therapists differ in their effectiveness, that these differences are small (depending on the study design) and that they seem – at least in naturalistic samples – to be reliable. This situation is to be expected, since, in a naturalistic situation, variability in therapist skills would be a natural phenomenon. Besides these relatively homogeneous findings in meta-analyses, the estimated proportion of variance that is attributable to therapists varies enormously between individual studies and samples. This becomes obvious in the meta-analysis reported by Baldwin and Imel (2013), where the estimated proportion of variance that is attributable to therapists varies between 0% and 50%. Research has not focused on the reasons for this variability in TEs across naturalistic studies. However, it might partly be explained by small sample sizes leading to distorted results. The question remains as to how much sample size issues contribute to this heterogeneity in comparison to real variations in outcomes between therapists.

Since the emergence of multilevel modeling (MLM), it has become the standard method for investigating TEs (e.g. Adelson & Owen, 2012; Okiishi et al 2006). This method, which models the hierarchical structure of the data, with patients 'nested' within therapists, derives a TE that corresponds to the intraclass correlation coefficient (ICC) (see Raudenbush & Bryk, 2002). Hence, the accuracy and reliability of model parameter estimates and therefore the robustness of TEs, depends on the sample size. In the standard two-level multilevel model, three sample size parameters are relevant: the number of patients (Level 1), the number of therapists (Level 2) and the number of patients treated by each therapist. Because of these three different parameters, it becomes clear that sample size calculations that have been developed for traditional single-level designs cannot be applied to MLM.

Studies of sample size for cluster randomized trials (CRTs), where groups of subjects, rather than individuals, are randomized (Eldridge, Ashby, & Kerry, 2006), have recognized the problem of ignoring the hierarchical structure and the 'group effect' (i.e. the elevated risk of type 2 errors). In response, methods and formulas have been developed that take into the

'group effect' when making sample size calculations (e.g. Gao, Earnest, Matchar, Campbell, & Machin, 2015; Moerbeek, 2014; Shoukri, 2004). However, these methods rely on a reliable a priori estimate of the group effect which in psychological therapies, given the heterogeneity of TEs above, is uncertain and still open to discussion. One example as to how an inadequate sample size can result in very different TEs is the reanalysis of the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP; Elkin et al., 1989). This study was originally designed to investigate the effectiveness of two forms of brief psychotherapy (cognitive behavior therapy and interpersonal psychotherapy) in comparison to a pharmacotherapy and placebo condition. The sample contained 17 therapists who treated between 4-11 patients each. Using the same sample, Elkin, Falconnier, Martinovich and Mahoneya (2006) could not find variance associated with therapists, whereas Kim and colleagues (2006) identified a TE of approximately 8%. The small sample size, along with other issues, has been identified as a cause of these contrary results (Crits-Christoph & Gallop, 2006; Elkin, Falconnier, & Martinovich, 2007; Lutz & Barkham, 2015; Wampold & Bolt, 2006).

To date, sample size issues of MLM have been approached via simulation studies that result in formulating different guidelines regarding minimum sample size. Some researchers suggest a minimum sample size of 30 groups on level 2 (therapists) and 30 units per group on level 1 (patients) to have enough power in a two-level design (Kreft, 1996). Maas and Hox (2005) argue that a major restriction in MLM is higher-level sample size. In a simulation study, they showed that samples of 100 generic clustering units led to unbiased estimates of variance components and standard errors. In their study, a large number of level 2 units appeared to be more important than the number of units on level 1. The lowest group size included in the simulation analyses was 5 units on level 1, which resulted in unbiased estimations if enough units on Level 2 were included in the samples.

Some researchers incorporate an alternative perspective that draws attention to the focus of the research question (Hox, 2010; Raudenbush & Bryk, 2002). If the investigation aims to analyze random effects, Hox (2010) recommends applying a 100/10 rule: 100 therapists on level 2 and a group size of 10 patients per therapist resulting in a sample of 1000 cases. If cross-level interactions are of interest, the equivalent recommendation is a 50/20 rule: 50 therapists treating 20 patients each, which results, again, in a sample of 1000 cases. Other research has focused on the power within three-level longitudinal models with repeated measures on level 1, patients on level 2 and therapist on level 3. Based on a simulation study, de Jong, Moerbeek and van der Leeden (2010) provide recommendations concerning different sample size combinations for all three levels to reach a power of 0.80.

In summary, the above mentioned simulation studies supply researchers with inconsistent rules of thumb with relatively high average sample sizes on each level. So far, very few research studies have been able to realize these sample size demands (e.g. Saxon & Barkham, 2012). Nonetheless, several studies have at least approximately reached tolerable sample sizes (e.g. Dinger, Strack, Leichsenring, Wilmers, & Schauenburg, 2008; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Nielsen, & Ogles, 2003) with an average of 55 therapists per dataset, who treated at least 10 patients each, resulting in samples ranging from N = 1,779 to N = 2,554 cases. However, in Baldwin and Imel's (2013) review, 43 out of 46 studies can be classified as having serious sample size problems. The median number of therapists within these studies was 9 with a median of 7.6 patients per therapist. In contrast to these "real-world" findings, a simulation study by Musca and colleagues (2011) did not even include groups smaller than ten cases. However, in naturalistic samples it is common to have therapists with fewer than ten treated patients (see Baldwin & Imel, 2013).

Due to the reported variability in the size of TEs and the apparent influence of sample sizes, the main aim of the present study was to develop empirical estimates of TEs for varying sample size conditions and to explore sample size factors, which may affect their magnitude

as well as their stability. First, we individually examined eight naturalistic datasets regarding the extent of TEs, while controlling for initial impairment in therapist caseloads. In line with the existing literature, we expected to find substantial differences in TEs between datasets, but with all of them showing significant TEs. After standardization and integration into one sample, we anticipated finding an average significant TE of about 5%.

Second, we developed sample size tables for future research via resampling. The aim was to provide practical information to aid the planning of future studies in this field and to complement simulation work on providing sample size guidelines in multilevel analyses of TEs.

Method

Original Datasets

The study sample included eight datasets drawn from 3 countries (US, UK and Germany), including 6 different outcome measures routinely collected between 1990 and 2013 and cumulating in aggregated data from 48,648 cases treated by 1,800 therapists. All individual datasets complied with local ethics committee approvals where necessary. In the following section, the eight international samples are described individually.

The *University Outpatient Clinic* sample from southwestern Germany comprised 668 psychotherapy outpatients and 97 therapists who each saw between 2 and 18 patients (M = 8.78, SD = 3.70). Therapists were all part of a Cognitive Behavioral Therapy (CBT) based post-graduate training program. Patients attended between 3 and 98 sessions (M = 33.46, SD = 17.31). The patients' mean age was 36.35 (SD = 12.49; range = 15-74); 70.3% were women; 40.1% had a primary diagnosis of major depressive disorder, 18.7% were diagnosed with anxiety disorder, 16.6% had an acute stress and adjustment disorder, 5.8% had a dysthymic disorder, 4.0% an eating disorder, 1.2% were diagnosed with a personality disorder, and 13.4% were classified with another psychological disorder. The Brief Symptom Inventory (BSI; Franke, 2000) was used as the primary outcome measure.

The *Techniker Krankenkasse* sample was based on a health insurance pilot project that investigated quality management in outpatient psychotherapy in Germany between 2005 and 2010, supported by the German health insurance company *Techniker Krankenkasse* (TK; Lutz, Böhnke, & Köck, 2011). A subsample of the TK-project was used in this paper. It comprised 636 psychotherapy outpatients and 120 therapists who saw between 2 and 18 patients each (M = 8.31, SD = 4.94). Therapists were from different theoretical orientations: 69.8% had a CBT background, 34.9% were trained in psychodynamic psychotherapy, whereas 3.1% had a psychoanalytic orientation (multiple answers possible). Patients attended between 5 and 143 sessions (M = 35.66, SD = 20.86). The patients' mean age was 45.06 (SD = 11.30; range = 21-77); 68.2% were women and 97.2% were German. 38% had a major depressive disorder, 21.2% were diagnosed with an acute stress and adjustment disorder, 19.2% had an anxiety disorder, 7.1% had a dysthymic disorder, 2.4% were diagnosed with an eating disorder, 2.2% with a personality disorder and 10% were classified with another psychological disorder. For the TK project, the BSI was also one of the primary outcome measures (Franke, 2000).

The *University Outpatient Clinic in Midwest Germany* sample comprised 752 patients treated by 71 therapists. Therapists were either trained or part of a post-graduate training program with CBT as their theoretical orientation. Each therapist treated between 2 and 26 patients (M = 13.02, SD = 4.91). Patients attended between 4 and 90 sessions (M = 30.62, SD = 17.72). The patients' mean age was 37.29 (SD = 11.71; range = 16-74); 56.9% were women; 44.9% were diagnosed with an anxiety disorder, 22.1% had a major depressive episode, 8% had an acute stress and adjustment disorder, 6.6% were diagnosed with an eating disorder, 3.4% had a dysthymic disorder, 1.9% were diagnosed with a personality disorder, and 12.9% were classified with another psychological disorder. Like the other German samples, the BSI (Franke, 2000) was used as the primary outcome measure in this dataset.

The *CelestHealth* dataset was based on data from 26 centers comprising 20 college counseling centers, four primary care medical centers, and two private mental health centers located in the US. The sample comprised 11,356 patients treated by 401 therapists. Each therapist treated between 2 and 203 patients (M = 63.74, SD = 43.94). Therapists included psychologists, psychiatrists, clinical social workers, and trainees, all reflecting a varied professional background and theoretical orientation. Furthermore, treatment duration was variable and not subject to strict time limits so that patients attended between 3 and 154 sessions (M = 8.66, SD = 8.90). All patients were older than 18 years and a majority were female (63.5%). No information on diagnosis was available for this sample. The primary outcome measure was the *Behavioral Health Measure-20* (BHM-20; Kopta & Lowry, 2002).

The *Compass Tracking System*, originally called *Integra Outpatient Treatment Assessment system* (IOTA; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lueger et al., 2001; Lyons, Howard, O'Mahoney, & Lish, 1997), is a quality monitoring system and one of a number of comprehensive assessment batteries that has been used to measure progress in outpatient mental health. The dataset gathered with the assistance of the Compass System comprised 1,194 psychotherapy outpatients who were treated by 60 therapists in different settings in the US (Lutz et al., 2007). Therapists were part of the national provider network of an American managed care company. All therapists had formal training and at least 1 year post qualification experience. They varied in professional background and theoretical orientation that was not systematically recorded. Each therapist treated between 10 and 77 patients (M = 28.79, SD = 19.50). Treatment duration was not subject to strict limits so that patients attended between 3 and 120 sessions (M = 9.60; SD = 10.49). The patients' mean age was 36.40 (SD = 9.50); 73% were women; 59% were married, 24% were single, and 18% were separated, divorced, or widowed; 43.9% were diagnosed with an affective disorder, 28.4% had an acute stress and adjustment disorder, 8.8% had an anxiety disorder, 0.8% were diagnosed with an eating disorder, 4.7% had another psychological disorder and for 13.4% of

the cases, the diagnosis was missing. The primary outcome measure of the Compass Tracking System was the Mental Health Index (MHI; Howard, Brill, Lueger, O'Mahoney, & Grissom, 1993b).

The *University Counseling Center* dataset was collected at a large site in the US. It comprised 2,561 patients treated by 143 therapists. All of the therapists were doctoral level students in training or doctoral licensed mental health professionals. They had a variety of treatment orientations, with most integrating two or more theoretical systems (e.g. cognitive and behavioral). Each therapist treated between 2 and 155 patients (M = 56.30, SD = 47.38). Patients attended between 3 and 102 sessions (M = 8.50; SD = 8.21). The patients' mean age was 31.84 (SD = 5.12; range = 21-74); 58.6% were women and 91.9% were American. 17.3% were diagnosed with an affective disorder, for 7.7% the diagnosis was deferred, 7.5% had an acute stress and adjustment disorder, 5.9% were diagnosed with an anxiety disorder, 2.6% had an eating disorder, 0.3% were diagnosed with a personality disorder, 5.2% had another psychological disorder, whereas 31.2% had no psychological disorder and 22.3% no diagnosis at all. The primary outcome measure was the Outcome Questionnaire-45 (OQ-45; Lambert, 2004).

The *Clinical Outcomes in Routine Evaluation (CORE) Practice-Based Evidence National Database-2008* comprised 25,842 patients treated by 789 therapists in counseling and psychotherapy centers in the United Kingdom between 1999 and 2008. All therapists had training in psychological therapy and at least 1 year post qualification experience. Furthermore, a variety of treatment approaches were offered, whereas none of the therapists consistently followed a formal manualized protocol. Each therapist treated between 2 and 400 patients (M = 103.31, SD = 87.21). Patients attended between 3 and 117 sessions (M = 6.83; SD = 4.37). The patients' mean age was 40.27 (SD = 11.93, range = 16-65); 71.3% were women. No formal diagnosis was recorded. Nevertheless, therapists identified patients' presenting problems. This indicated that 70.8% were experiencing depression, 42.1% at a

moderate to severe level, while 78.4% were experiencing anxiety, 55.5% at a moderate to severe level. The primary outcome measure of this sample was the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM; Barkham et al., 2001; Evans et al., 2002).

The *Improving Access to Psychological Therapies* (IAPT) dataset comprised 5,639 patients treated by 119 therapists and was collected in North England between 2008 and 2010. Therapists in this sample included qualified CBT practitioners delivering high intensity psychotherapy (up to 20 sessions), registered mental health nurses, counsellors, and psychological well-being practitioners (PWPs) delivering low intensity and brief (less than 8) CBT-oriented guided self-help interventions. Treatments were delivered in a stepped care system, with the majority of patients accessing brief interventions and progressing to high intensity psychotherapy if required, as recommended by English clinical guidelines (National Institute for Health and Care Excellence, 2011). Each therapist treated between 2 and 163 patients (M = 80.14, SD = 42.60). Patients attended between 3 and 21 sessions (M = 6.63; SD = 3.81). Patients' mean age was 39.06 (SD = 13.54, range = 16-98). The majority of patients were women (65.4%); 30.4% were diagnosed with an affective disorder, 22.9% had a mixed anxiety and depression disorder, 19% had an anxiety disorder, 2.2% were diagnosed an obsessive-compulsive disorder, 1.4% had an acute stress and adjustment disorder, 0.8% had an eating disorder, 23.2% had another psychological disorder, and 22.3% no diagnosis at all. The relevant outcome measure in the IAPT dataset was the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001), self-completed by patients on a session-by-session basis.

Instruments

Brief Symptom Inventory (BSI; Franke, 2000; German translation of Derogatis, 1975). The BSI is a 53-item self-report symptom inventory for the evaluation of physical and psychological symptoms within the last week. It is the brief form of the Symptom Checklist-

90-R (SCL-90-R; Derogatis, 1977). The instrument taps 9 primary dimensions: *somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation* and *psychoticism*. In this study, only the *Global Severity Index* (GSI) was calculated by averaging all items. The items are scored on a 5-point Likert scale ranging from 0 (*not at all*) to 4 (*extremely*). The internal consistency of the BSI has been found to be α = .92 and the retest-reliability r_{tt} = .90 (Franke, 2000).

Behavioral Health Measure-20 (BHM-20; Kopta & Lowry, 2002). The BHM-20 is a 20-item self-report questionnaire for the evaluation of mental health. The instrument comprises three subscales: *well-being*, *psychological symptoms* and *life functioning*. The *Global Mental Health Index* (GMH) was used for the present paper, which is calculated by averaging the 20 items. Clients rate the items on a Likert scale ranging from 0 (*extreme distress/ poor functioning*) to 4 (*no distress/ excellent functioning*). The scales were adjusted so that higher scores indicated more psychological distress. The internal consistency of the BHM has been found to be α = .89 to .90 and the retest-reliability r_{tt} = .80 (Kopta & Lowry, 2002).

Mental Health Index (MHI; Howard et al., 1993b). Within the Compass Tracking System, a patient's progress in outpatient treatment was measured based on three scales capturing both their own as well as the clinician's perspective (Howard et al., 1993b). The present study focused on the scales capturing the patient's perspective, which comprised 68 items. The three scales *subjective well-being*, *current symptoms* and *current life functioning* are in line with the three phases of the phase theory of psychotherapy: *remoralization, remediation* and *rehabilitation* (Howard, Lueger, Maling, & Martinovich, 1993a). The three scales are combined into a Mental Health Index (MHI) that was used in the current analyses with higher scales indicating more psychological distress. The internal consistency of the MHI has been found to be α =.88 and the test-retest correlation r_{tt} = .82 (Howard et al., 1993b). The scales were adjusted so that higher scores indicated more psychological distress.

Outcome Questionnaire-45 (OQ-45; Lambert, 2004). The OQ-45 is a self-report instrument that captures mental health functioning over the course of the last week. The questionnaire can be administered at the beginning as well as over the course of treatment to track and measure client progress in psychotherapy. The 45 items are scored on a five-point Likert scale ranging from 0 (*never*) to 4 (*almost always*), resulting in a range of possible scores from 0 to 180. Besides the global sum score, the OQ-45 comprises three subscales: *symptom distress, interpersonal functioning* and *social role functioning*. In this study, the total score was utilized so that higher scores indicated more symptom severity. Internal consistency reliabilities have been found to vary from $\alpha = .70$ to .93 for the total scale and subscales. Test-retest reliabilities range from $r_{tt} = .78$ to .84 (Lambert, 2004; Lambert et al., 1996).

Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM; Barkham et al., 2001; Evans et al., 2002). The CORE-OM is a self-report measure comprising 34 items addressing four different domains: *well-being, symptoms, functioning* and *risk.* Items are scored on a 5-point Likert scale from 0 to 4 anchored by the following terms: *not at all, only occasionally, sometimes, often* and *all or most of the time*. A global score is calculated as the mean of all completed items multiplied by 10, yielding a range from 0 to 40 with higher scores indicating more symptom severity. The internal consistency of the CORE-OM has been found to range from $\alpha = .93$ to .95 with a test-retest reliability of $r_{tt} = .90$ (Barkham et al., 2001; Evans et al., 2002).

Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001). The PHQ-9 comprises items drawn from the primary care evaluation of mental disorders (PRIME-MD), which has been validated for use in primary care. The 9-item depression scale used in this paper captures depression corresponding with DSM-IV criteria as well as general symptom severity. Items are rated on a scale ranging from 0 (*not at all*) to 3 (*nearly every day*). For the purpose of this paper, the global sum score was calculated ranging from 0 to 27. The internal reliability of the PHQ-9 has been found to be $\alpha = .89$ and its validity has been shown in a

variety of settings and populations (Kroenke et al., 2001; Manea, Gilbody, & McMillan, 2012).

Data prescreening and standardization

All data included in the analyses were prescreened concerning the following criteria: a) individual patient data comprised pre- and post-therapy measures on the appropriate outcome instrument; b) a unique ID was available for each therapist; c) each therapist treated a minimum of two patients; and d) each case comprised at least 3 sessions.

Moreover, as described previously, six different instruments were used to assess outcome across the eight samples. For this reason, a standardization procedure was necessary to integrate the subsamples into one large dataset. The most common method for standardization is to perform a z-transformation, where the sample mean is subtracted from each score and the difference is divided by the sample standard deviation. Although the eight samples were routinely collected, the level of patient impairment cannot be presumed to be equal across institutions, datasets, and countries. A normal z-transformation would not take these distinctions into account and, moreover, this procedure might confound the size of the TE. We reasoned that standardizing each individual dataset on the mean and standard deviation of an appropriate measure-specific outpatient reference sample drawn from current psychotherapy research would obviate this potential confound. Hence, for each of the six instruments, the mean and standard deviation of a clinically impaired population were identified. Using the resulting reference values, the pre and post scores of the associated datasets were standardized. Subsequently, all eight datasets were integrated into one large dataset that represented the basis for the following analyses.

Data Analytic Strategy

All eight samples contained a hierarchical data structure, for which multilevel modeling (MLM) has been established as the method of choice (Hox, 2010; Raudenbush & Bryk, 2002). To analyse the TE and its variation in each of the eight samples, two-level

models were calculated with patients at level 1 and therapists at level 2 (equations are reported in the Appendix). The two-level model partitions the total variability into two components: variance within patients at level 1 and between therapists at level 2. The variance associated with level 2 divided through the total variance is the TE (Baldwin & Imel, 2013). All models that were used to calculate the TE included pre-treatment intake scores on the relevant outcome measure to control for individual differences in pre-test levels. This variable was standardized as described above and therefore all TEs were estimated for the average initial patient severity. In all models, division of level 2 variance through total variance resulted in intraclass correlation (ICC), which is a synonym for the TE (Hox, 2010). The higher the ICC, the larger the differences between therapists concerning the outcome variable of interest: patient outcome. Furthermore, we tested the possibility of a random slope model for all eight datasets, where the relationship between pre-treatment scores and outcome was allowed to vary between therapists. The Akaike information criterion (AIC) was used to investigate which model fit the data best, whereas smaller values indicate a better model fit (Hox, 2010)[1].

In a next step, after integrating the eight samples into one dataset, a three-level hierarchical model was conducted with patients at level 1, therapists at level 2 and datasets at level 3 (equations are reported in the Appendix). The three-level model partitions the total variability in outcome into three components: variance within patients on level 1 (σ^2), between therapists on level 2 (τ_π), and between datasets on level 3 (τ_β). As in the two-level model, the variance associated with level 2 yields the TE and was calculated using the ICC corrected for initial patient severity. Again, a random slope model was considered, where, once more, the AIC served as the fit criterion. The variance associated with level 3 represents the dataset effect. Although eight units at level three is not sufficient to reliably interpret the

[1] To reduce the complexity of this paper, AIC values are not reported in detail but can be requested from the first author.

dataset effect, we included the third level because the analyses of the eight individual datasets revealed a large variation in TEs. By including the third level in the model, these dataset differences could be extracted from the total variance, allowing the estimate of the TE to become more precise.

Investigation of sample size issues. The investigation of sample size issues in relation to the extent and reliability of TEs was achieved by examining different sample size conditions. A basic subsample was formed comprising only therapists who treated a minimum of 30 patients. This resulted in a core subsample of 484 therapists (patient N=36,263). In reducing *the number of therapists* and the *number of patients per therapist*, different sample size conditions were produced. For each sample size condition, 1,000 samples were randomly selected out of the existing core subsample. This allowed us to estimate the mean TE across 1,000 samples for each sample size condition. Furthermore, confidence intervals (CIs) were computed, which were used as indices of precision for the estimated mean TEs. The reference for the width of the CIs was based on the results of the two existing meta-analyses in this research field. These two studies estimated TEs between 5% and 9% resulting in a range of 4% (Baldwin & Imel, 2013; Crits-Christoph et al., 1991). On this basis, we decided to allow the CIs in our study to have a maximal range of 4%.

We reduced the *number of therapists* in intervals of 100 (400, 300, 200 & 100) and then from 50 the number of therapists decreased in intervals of 10 (50, 40, 30, 20 & 10). As soon as the number of therapists per dataset was reduced to 10, the reduction scheme changed and specified only 5 and then 2 therapists per dataset. In line with this, the *number of patients per therapist* was reduced starting with only those therapists that treated 30 patients. First, the reduction was implemented at intervals of 5 (30, 25, 20, 15, 10 & 5) and then in single steps (5, 4 & 3). Finally, sample size tables were generated that included information about mean TEs and CIs for each sample size condition. A total of 72 sample size conditions were computed.

As a result of the resampling procedure described above, two new variables were generated: *mean TE per sample size condition* and its *CI*. Each of the two variables served as an outcome variable in a multiple regression analysis that was conducted to investigate the influence of *number of patients per therapist* as well as *number of therapists per dataset*. All data analyses were conducted with the free software environment R version 3.1.0 (R Development Core Team, 2014). For MLM, the package lme4 was used (Bates, Maechler, & Bolker, 2013) whereas parameters were estimated via maximum-likelihood (ML) and p-values were calculated using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2014).

Results

Variability across datasets

The AIC revealed that a random intercept model had a better fit in the analyses of all eight datasets than a single level regression model, thereby indicating significant variability between therapists (level 2) even after adjusting for initial patient impairment. The two-level analyses revealed variability in TEs and effect sizes between the eight datasets (Table 1). TEs varied between 2.7% (IAPT dataset) and 10.2% (CORE Practice-Based Evidence National Database 2008) whereas effect sizes ranged between .49 (University Counseling Center in the UK) and 1.45 (CORE Pratice-Based Evidence National Database 2008). These heterogeneous results must be interpreted with care, based on the knowledge of dataset differences concerning treatment process (Complete vs. non-completer; see Table 1). Averaging the eight individual TEs led to a mean TE of 5.7%[2]. Furthermore, the random slope model improved model fit in all samples regarding the fit criterion AIC. This suggested significant variability between therapists concerning the relationship between pre-treatment scores and outcome.

[2] Because of the nested data structure, three sample size parameters must be considered when weighting the arithmetic mean: 1) number of patients 2) number of therapists 3) number of patients per therapist. The mean TE weighted for the number of patients is 7.2%, the mean TE weighted for the number of therapists is 7.1% and 5.75% if the mean TE is weighted by the mean number of patients per therapist.

Additionally, therapist variation of baseline estimates was investigated. Across the eight datasets a mean between therapist variation of 4.3% was detected. It ranged from 0.4% in the University Outpatient Clinic sample from southwestern Germany to 11.3% in the German TK sample.

Three level hierarchical analyses for the total dataset

The results for the aggregated dataset are displayed in Table 2. Initial patient impairment was a significant predictor and explained 25.3% of the variation in outcome. Dividing level 2 variance (therapist variation) by the total variance in model 1 led to a significant TE of 6.7%. Thus, most of the variation in outcomes (87.1%) was at the individual patient level (level 1). Again, including a random slope improved model fit regarding the AIC, suggesting that there were considerable differences between therapists regarding the relationship between initial impairment and treatment outcome. Comparing the residuals and 95%-CIs of each therapist with the average therapist outcome, we identified which therapists were above or below that average. This resulted in 225 therapists (12.5%) out of 1,800 who were identified to be above average in terms of the outcomes of their patients and 11.8% (N = 212) of the therapists who were below average. Consequently, 1,363 therapists (75.7%) were ranked as average regarding the outcome of their patients and could not be reliably differentiated from each other.

Investigating sample size issues

The results of the resampling procedure are presented in Table 3[3]. For each of the 72 sample size conditions, the mean TE as well as its CI were calculated within three-level hierarchical models allowing slope and intercept to vary between therapists (see Appendix). The mean TEs and associated CIs were used as outcome variables in two individual multiple regression analyses that were run to evaluate the impact of sample size parameters. First, the

[3] Due to shortage of space and clarity, not all sample size conditions are displayed in Table 3.

influence of *number of patients per therapist* and *therapists per dataset* as well as their interaction on the mean TE were computed. The two predictor variables and their interaction were significant, $F(3, 135) = 33.90$, $p < .001$ (Table 4) and explained 43.5%[4] of the variance in TEs. A second multiple regression was conducted to predict the range of the CIs. Again, covariates were significant $F(3, 135) = 50.99$, $p < .001$ and explained 53.7%[5] of the variation in the range of CIs.

Visual representations of the resampling procedure are given in Figures 1 and 2. These figures illustrate the influence of the two sample size parameters on the magnitude of TEs (Figure 1) and the width of CIs (Figure 2).

Application

The aim of these analyses was to provide researchers with guidance on sample sizes for an accurate estimation of TEs. We suggest interpreting the results in two consecutive steps. First, researchers should start with Figure 1, which depicts the mean TE for each sample size condition. As reference for an empirical TE, we used the 6.7% TE from the present study. After deciding on a sample size that meets the reference TE, a researcher should check if this sample is sufficient to result in a reliable TE. Second, in Figure 2 the width of the CIs per sample size condition are presented, which can be used as a measure for the precision of the estimation. The smaller the differences between the upper and the lower bound of the CI, the more reliable the computed TE.

Assume that in the planning phase of a naturalistic study, the aim is to investigate the TE in an outpatient clinic. A total of 10 therapists have been recruited to join the study and the question is: are 10 therapists sufficient to precisely estimate TEs? In the first step, Figure 1 indicates that each of these 10 therapists needs to treat at least 10 patients to reach the

[4,5] The explained variance (R^2) was calculated in accordance with the recommendations of Hox (2010).

reference TE. After deciding on a sample size that meets the reference TE (see Figure 1), a researcher should also check if this sample is sufficient to result in a reliable TE (see Figure 2). In this case, Figure 2 indicates that 10 therapists at level 2 is not recommended, as the CI exceeds 4% (CI difference = 12.27%; Table 3), thereby yielding an unreliable estimation. In this case, the number of patients per therapist cannot compensate for the small number of therapists at level 2. Hence, the study requires a further 30 therapists ($N_{therapist} = 40$). Additionally, the number of patients per therapist must be increased to 30 in order to reach a sample size that more precisely estimates the TE (Figure 2). This example indicates that minimum sample sizes at both levels are necessary. But given the minimum sample size at each level, Figure 1 and Figure 2 also show that sample size limits at one level can be partially compensated by those at the other level.

In the context of minimum sample sizes on both levels, Figure 1 shows that a sample to investigate TEs should have at least 4 patients per therapist. With smaller group sizes than 4 the TE will be overestimated, although the CI is within the reference range (Figure 2). It should be noted that with a group size of 4 cases, the sample needs to include at least 300 therapists, yielding a sample of 1,200 patients. With regard to level 2, at least 40 therapists per sample are needed in order to be able to estimate the TE reliably. Again, it should be highlighted that in a sample with 40 therapists, each therapist needs to treat at least 30 cases thereby leading to a sample of 1,200 patients.

Discussion

Twenty-five years ago, Kazdin and Bass (1989) raised the question concerning the extent to which comparative outcome studies are adequately powered in the field of psychotherapy outcome research. Their findings suggested that most of the studies that compared alternative treatments were insufficiently powered to detect small-to-medium effect sizes. However, at the time, investigations of large naturalistic datasets were rare. By contrast, the collection and investigation of large datasets is increasingly commonplace in current

research communities. In response, we consider that the question of statistical power and reliability of estimates can be raised in a different context regarding studies investigating TEs in practice. The question we examined is whether sample sizes of psychotherapy outcome studies are sufficiently large to reliably detect differences between therapists, if they exist.

Estimates of sample sizes required in studies of TEs have been addressed using simulation studies (e.g. Bell, Morgan, Schoeneberger, Kromrey, & Ferron 2014; Maas & Hox, 2005). These studies have delivered inconsistent results and corresponding rules for sample sizes at each level of the MLMs. In addition, this guidance often does not reflect the sample structure in research studies. In view of the above, the present study is the first empirical investigation of sample size issues focusing on MLM and TEs in the context of naturalistic study designs. To date, no study has estimated the TE in a naturalistic sample with such a large sample size (N = 48,648). This is important, considering its implications for interpreting the percent of variance in outcome that can be attributed to patients and treatments.

Practical sample size tables for calculating TEs were the result of the investigation of eight naturalistic datasets and resampling procedures. These can be utilized to identify minimum samples sizes in future practice-oriented studies focusing on TEs. The values in the tables reveal that there is a degree of flexibility in the numbers of therapists and patients per therapist required, depending on the approximated CI. For example, a variable number of therapists and patients per therapist is possible as long as an overall sample size of 1,200 patients is achieved, which allows for an estimated TE within a CI lower or equal to 4%. This means that at least 4 patients per therapist with 300 therapists or at least 40 therapists treating 30 patients are necessary to render sufficiently accurate parameter estimates. This number is consistent with the existing literature on simulation studies using MLM, which suggest an overall sample size of approximately 1,000 cases (e.g. Hox, 2010; Raudenbush & Bryk, 2002). The sample size of 1,200 cases highlights the limitations that occur within small services that try to analyse TEs. Prospectively, nation-wide services and systems rather than

small services will provide sufficiently large datasets to overcome sample size limitations. One example is the IAPT program in England, which is a government funded initiative to offer patients routine psychological treatment. Within this program, data is collected and merged from all affiliated services, resulting in a large and expanding database, which also provided a dataset for the current analyses. More such national practice networks would advance research possibilities in the context of TEs.

Crucially, the present study also investigated the consequences of samples comprising small numbers of therapists as well as samples with small numbers of patients per therapist. Attempts to bridge the scientist-practitioner gap are hindered where the research demands placed on routine services are unrealistic. Hence, sample size guidelines were formulated for small numbers on level 1 (patients per therapist), which seem to be realistic to obtain in routine care datasets. Results are displayed in easy-to-read sample size tables, which can be flexibly applied by researchers to evaluate the appropriateness of their assessment structure, if TEs are considered. Furthermore, the tables can be used to get a rough estimate of the potential CI related to existing samples. This can help to plan studies and/or to understand the heterogeneity in results between different studies. Obviously, guidelines for studying therapist effects drawn from routine care have implications for studying therapist effects in clinical trial research (recall the controversy ignited by the re-analysis of the NIMH TDCRP dataset discussed in the introduction). Given the small number of therapists and small number of patients per therapist in some clinical trials, therapist effects in such studies should be interpreted with caution.

Furthermore, applying a multiple regression approach, we analyzed the impact of the number of patients per therapist as well as the number of therapists as predictor variables using TEs and their CIs as outcome variables. The two sample size variables and their interaction explained approximately 50% of the variance in TEs and its CIs, making practice-oriented sample size guidelines even more important. Moreover, the results suggest that

different sample sizes between studies might be one important source of the observed heterogeneity in this research field. In line with these findings, the investigation of the eight naturalistic datasets yielded TEs ranging from 2.7% - 10.2%. Interestingly, this range is comparable with the existing literature, where some studies have reported TEs near zero (e.g. Ehlers et al., 2013; Elkin et al., 2006; Owen, Tao, & Rodolfa, 2010) while other studies have reported TEs of 10% or higher (e.g. Boswell, Castonguay, & Wassermann, 2010).

Besides varying sample sizes, the descriptive heterogeneity of the datasets in this article might deliver an additional explanation for the inconsistent results regarding the size of TEs. This extended range held even when a standardization procedure based on the average impairment of a clinical reference sample was implemented to control for the impact of initial impairment. Initial patient severity was found to be a significant predictor in all naturalistic datasets as well as in the integrated total dataset, which replicates former research (Saxon & Barkham, 2012). Additionally, in all eight datasets we consistently found the random slope model to be significantly superior as compared to the fixed slope model. This result suggests that there are differences between therapists in all datasets regarding the relationship between pre-treatment and post-treatment scores, indicating variability between therapists in terms of how much intake severity impacts treatment outcome. The random slope model showed the best model fit in the three-level MLM, which further supports this interpretation. At the moment, we do not know why some therapists seem to be similarly effective, no matter how impaired patients are and why others are less effective in adapting to different initial impairment levels. This question warrants further research and has implications for training as well as practice.

After integrating the data into one large sample, our three-level MLM approach showed that about 6.7% of the variance in outcome was explained by therapist differences. This was slightly higher than the 5.7% TE that was calculated when simply averaging the TEs of the eight individual datasets. Hence, the hierarchical model enhanced the effect by

correcting for level 3 influences. In sum, the size of the TE in the three-level MLM was comparable to the findings in the most recent meta-analysis in this field, which suggested that approximately 7% of the variance in outcome was associated with therapists in naturalistic study designs (Baldwin & Imel, 2013). In addition, the distribution of therapist effectiveness was akin to that reported by Saxon and Barkham (2012), although our data revealed approximately 10% more therapists to be average.

The main limitation of the present study relates to the heterogeneity of the investigated samples and the outcome measures as well as the lack of consistent additional predictors that might explain variance associated with therapists. It is important to note that in the analyses, some datasets were considerably larger (CORE Practice-Based Evidence National Database 2008: N = 25,842) in comparison to others (TK-project: N = 363) and therefore contribute much more cases to the aggregated dataset. Therefore, we incorporated a third level in the MLM to correct for varying dataset influences. However, with only eight datasets available it was not possible to reliably estimate the impact of the third level or even to use predictors to try to explain dataset variance. More datasets would have allowed us to investigate the 'dataset effect' and potential further predictors (e.g. completer, non-completer, and number of sessions) of the heterogeneity in TEs besides sample size.

Although variations in sample sizes across datasets may partially explain the range of TEs in the current study sample, it is unlikely to be the only source of variability across datasets. Different clinical populations, therapists' backgrounds, intervention settings, case mix factors, etc. may have enhanced bias in the data. For example, the TE estimate in the IAPT dataset is very small despite an adequate sample size. Additionally, cultural differences as well as differences between countries' health care systems could have influenced bias. There are large differences between the US, UK and Germany concerning care systems, culture and therapist training. In summary, it must be mentioned that it was not possible to

control for all sources of variability that may have impinged TE estimates, not only in the investigation of individual datasets, but also in the integrated study sample.

In addition to differences in datasets, the variety of outcome instruments may have also contributed to the heterogeneity of TEs. Instruments may differ in their ability to capture important variations in outcome and could therefore lead to different TEs. For example, Huppert et al. (2001) analyzed data from the Multicenter Collarborative Study for Treatment of Panic Disorder and found TEs ranging from 1% to 18%, depending on the outcome measure in the field of anxiety disorders. Hence, there may be an impact of instruments on TE sizes. However, in the current sample, it was not possible to investigate this influence specifically, as measures were confounded with datasets as well as countries. Consequently, the results must be interpreted with care, based on the knowledge that datasets and instruments can hardly be disentangled. However, we incorporated a third level in the MLM to control for datasets. Furthermore, the standardization procedure allowed us to investigate the TE from an overall perspective. Nevertheless, further research should consider the impact of different instruments when interpreting and examining the heterogeneity of TEs between studies.

An additional limitation concerns the interpretation of sample size tables. The analyses were conducted for pre-post models, which were corrected for initial impairment. This model has emerged to be the most applied model in TE research (Baldwin & Imel, 2013; Saxon & Barkham, 2012). However, there is a broad range of more complex models in the field of multilevel modeling. Raudenbush and Bryk (2002) pointed out the more complex the model (e.g. more predictors), the larger the required sample sizes. In line with this, we must state that our results are limited to the pre-post model described above and that we cannot make any recommendations concerning more complex models. One example is growth curve models, which analyze nested longitudinal data with repeated patient measures on level 1 (e.g. Lutz et al., 2007). De Jong, Moerbeek and van der Leeden (2010) dealt with sample size issues

concerning these models within the evaluation of TEs. Moreover, results are constrained to maximum-likelihood estimations. Accordingly, other statistical approaches such as Generalized Estimating Equations (GEE) could have been used. However, in regard to the research question, multilevel modeling on the basis of maximum-likelihood estimations which focuses on the partition of variance associated with each level seemed to be the appropriate method (e.g. Burton, 1998; Gardiner et al., 2009). Furthermore, it is the most common approach for analyzing TEs (Baldwin & Imel, 2103).

Despite these limitations, this article provides researchers with real-world recommendations concerning sample sizes for optimal study designs when the aim is to analyse TEs in a pre-post design. In addition, CIs presented in this paper aid in the interpretation and evaluation of TEs within existing samples. Moreover, sample size tables provide researchers with a practical and easy to use tool for the future planning of studies examining TEs. As mentioned in the theoretical section, the accurate TE is still a subject of discussion in this research field. Accordingly, the paper is a contribution on the path of reaching consent. In this sense, the application of the paper could be to use the results as a priori estimates for analytic formulas to calculate sample sizes for future TE studies (for a review see Shoukri, 2004). In conclusion, the combination of sample sizes on each level is crucial for the accuracy of the investigation of TEs in practice-oriented research. Tables presenting different sample size scenarios might help researchers to improve study designs and to integrate the interpretation of results in this research area. There is much to be learned from studying therapists, whose treatment effects are well below or above average. Therefore, we encourage researchers to consider sample size as an important precursor to undertaking such analyses.

References

Adelson, L. J. & Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy, 49*(2), 152-162. doi: 10.1037/a0023990

Baldwin, S. A. & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.)*, Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258-297). New York, NY: John Wiley & Sons, Inc.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C.,…McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, 69, 184–196. doi:10.1037/0022-006X.69.2.184

Bates, D., Maechler, M. & Bolker, B. (2013). lme4: Linear mixed-effects models using s4classes [Software-Handbook]. (R package version 0.999999-2)

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology:* European Journal of Research Methods for the Behavioral and Social Sciences, 10(1), 1-11. doi: 10.1027/1614-2241/a000062

Beutler, L. E., Malik, A., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, W., & Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 227-306). New York, NY: John Wiley & Sons, Inc.

Boswell, J. F., Castonguay, L. G., & Wasserman, R. H. (2010). Effects of psychotherapy training and intervention use on session outcome. *Journal of Consulting and Clinical Psychology, 78*, 717-723. doi: 10.1037/a0020088

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to

account for correlated responses: An introduction to gear generalized estimating equations and multilevel mixed modelling. *Statistics in Medicine, 17,* 1261-1291.

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., ... Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research*, *1*, 81-91. doi: 10.1080/10503309112331335511

Crits-Christoph, P. & Gallop, R. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research, 16*, 178-181. doi: 10.1080/10503300500265025

Derogatis, L. R. (1975). *Brief Symptom Inventory.* Clinical Psychometric Research: Baltimore.

Derogatis, L. R. (1977). SCL-90-R: Administration, Scoring and Procedures Manual I. Clinical Psychometric Research: Baltimore.

De Jong, K., Moerbeek, M., & Van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects. *Psychotherapy Research, 20*(3), 273-284. doi: 0.1080/10503300903376320

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H. (2008). Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology, 64*, 344-354. doi: 10.1002/jclp.20443

Ehlers, A., Grey, N., Wild, J., Stott, R., Liness, S., Deale, A., ... & Clark, D. M. (2013). Implementation of cognitive therapy for PTSD in routine clinical care: effectiveness and moderators of outcome in a consecutive sample. *Behaviour Research and Therapy, 51*, 742-752. http://dx.doi.org/10.1016/j.brat.2013.08.006

Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International journal of epidemiology, 35*(5), 1292-1300. doi: 10.1093/ije/dyl129

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., . . . Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971-982. doi: 10.1001/archpsyc.1989.01810110013002

Elkin, I., Falconnier, L., Martinovich, Z. & Mahoneya, C. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16*, 144-160. doi: 10.1080/10503300500268540

Elkin, I., Falconnier, L. & Martinovich, Z. (2007). Misrepresentations in Wampold and Bolt's critique of Elkin, Falconnier, Martinovich, and Mahoney's study of therapist effects. *Psychotherapy Research, 17*, 253-256. doi: 10.1080/10503300601039816

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Toward a standardized brief outcome measure: Psychometric properties and utility of the CORE-OM. *The British Journal of Psychiatry, 180*, 51-60. doi: 10.1192/bjp.180.1.51

Franke, G. (2000). BSI: Brief Symptom Inventory von L.R. Derogatis (Kurzform der SCL-90-R) – Deutsche Version. Beltz Test GmbH.

Gao, F., Earnest, A., Matchar, D. B., Campbell, M. J., & Machin, D. (2015). Sample size calculations for the design of cluster randomized trials: A summary of methodology. Contemporary clinical trials, 42, 41-50.

Gardiner, J. C., Luo, Z., & Roman, L.A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine, 28*, 221-239. doi: 10.1002/sim.3478

Garfield, S. L. (1997). The therapist as a neglected variable in psychotherapy research. *Clinical Psychology: Science and Practice, 4*, 40-43. doi: 10.1111/j.1468-2850.1997.tb00097.x

Goldstein, H., Browne, W. & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*, 223-231. doi: 10.1207/S15328031US0104_02

Hofmann, S. G., & Barlow, D. H. (2014). Evidence-based psychological interventions and the common factors approach: the beginnings of a rapprochement?. *Psychotherapy, 5*, 510-513. http://dx.doi.org/10.1037/a0037045

Howard, K., Lueger, R., Maling, M., & Martinovich, Z. (1993a). A phase model of psychotherapy outcome: causal mediation of change. *Journal of Consulting and Clinical Psychology*, *61*, 678-685. doi: 10.1037/0022-006X.61.4.678

Howard, K., Brill, P., Lueger, R., O'Mahoney, M. & Grissom, G. (1993b). *Compass* outpatient tracking assessment: Psychometric properties. Integra.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist*, *51*, 1059-1064. http://dx.doi.org/10.1037/0003-066X.51.10.1059

Hox, J. (2010). Multilevel analysis: Techniques and applications (2. Aufl.). England: Routledge.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables, and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69*, 747-755. http://dx.doi.org/10.1037/0022-006X.69.5.747

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57*, 138-147. http://dx.doi.org/10.1037/0022-006X.57.1.138

Kim, D., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research*, *16*, 161-172. doi: 10.1080/10503300500264911

Kopta, S., & Lowry, J. (2002). Psychometric evaluation of the behavioral health

questionnaire-20: A brief instrument for assessing global mental health and the three phases of psychotherapy outcome. *Psychotherapy Research*, 12, 413-426. doi: 10.1093/ptr/12.4.413

Kreft, I. G. G. (1996). Are multilevel techniques necessary? An overview, including *imulation studies.* Unpublished manuscript. Los Angeles: University of California, Department of Statistics.

Kroenke, K., Spitzer, R.L., & Williams, J.B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*, 606-613. doi: 10.1046/j.1525-1497.2001.016009606.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random* and fixed effects for linear mixed effect models (lmer objects of lme4 package). (R package version 2.0-6)

Lambert, M. J. (1992). Psychotherapy outcome research: Implications for integrative and eclectic therapists. In J. C. Norcross and M. R. Goldfried (Eds.). *Handbook of Psychotherapy Integration*. New York: Basic Books.

Lambert, M. J. (2004). Administration and scoring manual for the OQ-45.2 (outcome *questionnaire)*. OQ Measures, LLC.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.2)*. Stevenson MD: American Professional Credentialing Services.

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. Journal of Consulting and Clinical Psychology, 69, 150-158. http://dx.doi.org/10.1037/0022-006X.69.2.150

Lutz, W., & Barkham, M. (2015). Therapist effects. *The encyclopedia of clinical psychology*. Blackwell-Wiley.

Lutz, W., Böhnke, J. R., & Köck, K. (2011). Lending an ear to feedback systems: evaluation of recovery and non-response in psychotherapy in a German outpatient setting. *Community Mental Health Journal*, *47*, 311-317. doi: 10.1007/s10597-010-9307-3

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54*, 32-39. doi: 10.1037/0022-0167.54.1.32

Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. (1997). *The measurement and* management of clinical outcomes in mental health services. New York, NY: John Wiley & Sons, Inc.

Maas, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92. doi: 10.1027/1614-1881.1.3.86

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off scores for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal*, *184* (3), E191-E196. doi: 10.1503/cmaj.110829

Moayyedi, P. (2004). Meta-analysis: Can we mix apples and oranges? *American Journal of Gastroenterology*, *99*, 2297-2301. doi:10.1111/j.1572-0241.2004.40948.x

Moerbeek, M. (2014). Multilevel modeling in the context of growth modeling. *Annals of Nutrition and Metabolism, 65*, 121-128. doi: 10.1159/000360485

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on Type-I error. *Frontiers in Psychology*, *2*, 1-6. doi: 10.3389/fpsyg.2011.00074

National Institute for Health and Care Excellence (2011). *Common mental health disorders: Identification and pathways to care.* [CG123]. London: National Institute for Health and Care Excellence. Retrieved on 21/04/2015 from http://www.nice.org.uk/guidance/CG123

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink:

An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy*, *10*, 361-373. doi: 10.1002/cpp.383

Owen, J., Tao, K., & Rodolfa, E. (2010). Microaggressions and women in short-term psychotherapy: Initial evidence. *The Counseling Psychologist, 38*, 923-946. doi: 0.1177/0011000010376093

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Ricks, D. F. (1974). Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In D. F. Ricks, M. Roff, & A. Thomas (Eds.), *Life history research in psychopathology* (Vol. 3, pp. 275-297). Minneapolis: University of Minnesota Press.

R Development Core Team (2014). R [Computer software]. Retrieved from http://www.R-project.org/

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, *80*, 535-546. doi: 10.1037/a0028898

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*, 251-271. doi: 10.1191/0962280204sm365ra

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine*, 21(24), 3757–3774. doi: 10.1002/sim.1330

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16*, 184-187. doi: 10.1080/10503300500265181

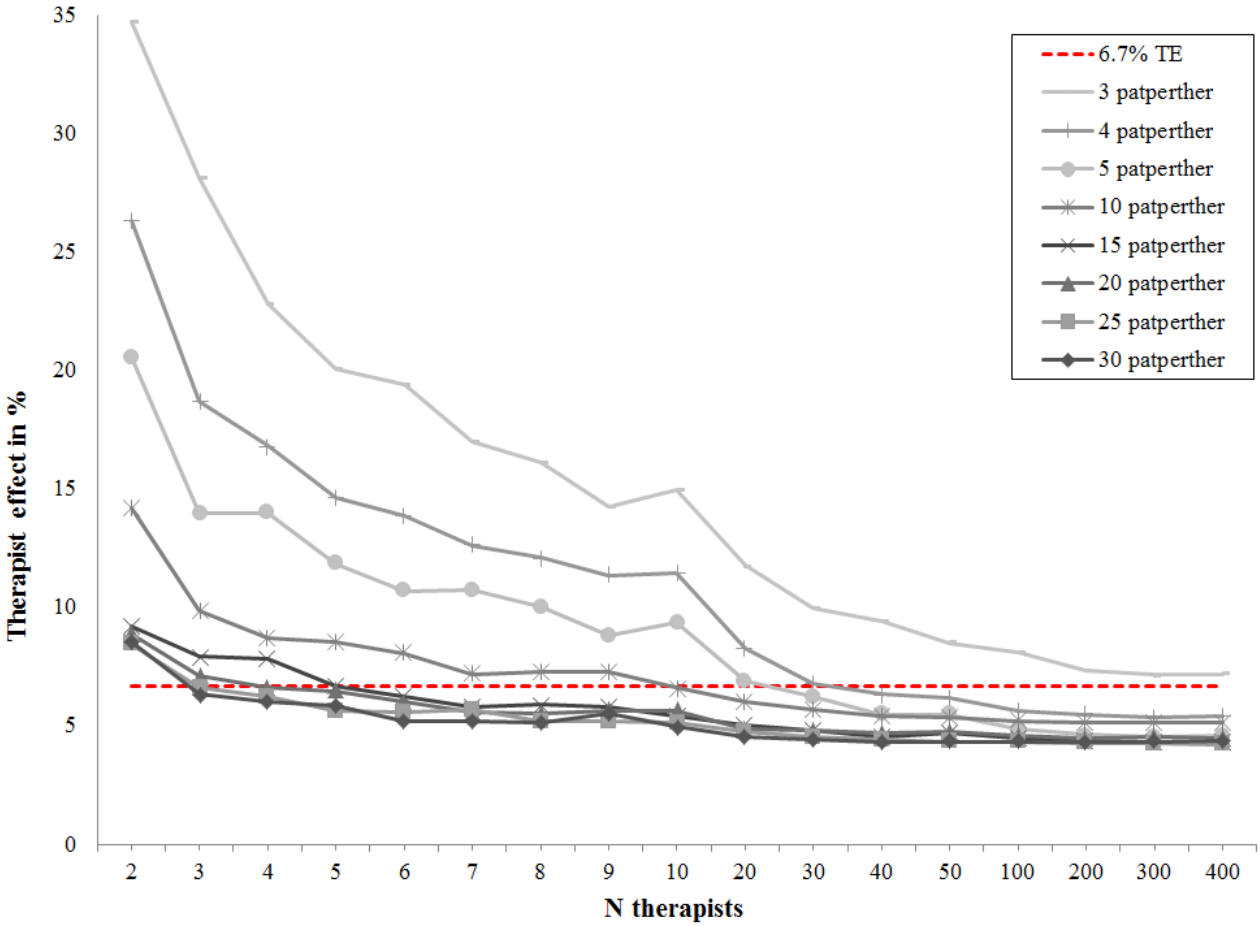

Figure 1. Influence of the group size (number of patients per therapist) and number of therapists on the estimated mean therapist effect of 1,000 samples. Note that the 6.7% therapist effect from the aggregated dataset was added as reference line in the graphic. Displayed results are based on a three-level model with random intercept and slope (see Appendix). patperther = number of patients per therapists; ICC = intraclass correlation/ therapist effect.

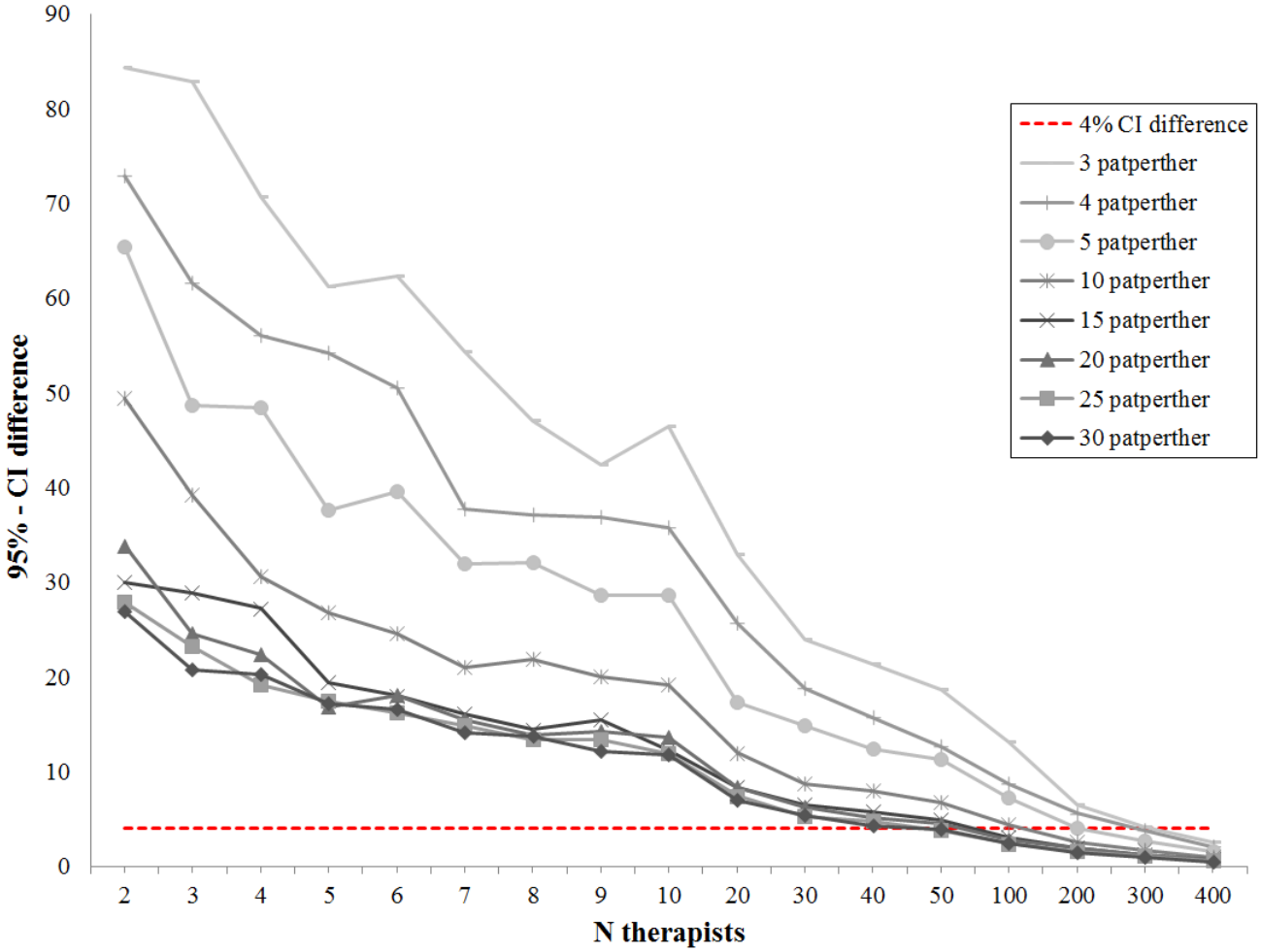

Figure 2. Influence of the group size (number of patients per therapist) and number of therapists on the size of the 95% CI of the estimated mean therapist effect of 1,000 samples. Note that 4% difference was added as reference line in the graphic. Displayed results are based on a three-level model with random intercept and slope (see Appendix). CI = confidence interval; patperther = number of patients per therapists

Table 1

Patient intake, outcome scores and therapist effects (TEs) for the eight naturalistic datasets.

Dataset	Country	Instrument	Intake		Outcome		Completer Sample	Number of sessions *M* (SD)	Effect size[10]	TE[11]
			M (SD)	Range	M (SD)	Range				
University Outpatient Clinic Southwest Germany[1]	GER	BSI	1.23 (0.67)	0.02–3.33	0.62 (0.55)	0–3.13	Yes	33.46 (17.31)[9]	.92	5.5%
TK-project[2]	GER	BSI	1.21 (0.66)	0.06–3.36	0.6 (0.53)	0–3.13	Yes	35.66 (20.86)	.94	9%
University Outpatient Clinic Midwest Germany[3]	GER	BSI	1.26 (0.72)	0.02–3.3	0.73 (0.65)	0–3.43	No	30.62 (17.72)	.73	5.5%
CelestHealth project[4]	US	BHM-20	2.55 (0.63)	0.2–4.0	2.94 (0.62)	0–4.0	No	8.66 (8.90)	.62	3.8%
Compass Tracking System[5]	US	MHI	48.08 (8.75)	22.96–77.21	54.15 (9.14)	22.31–77.50	No	9.60 (10.49)	.69	4.7%
University Counseling Center[6]	UK	OQ-45	65.06 (21.73)	6–128	54.36 (22.67)	0–150	No	8.50 (8.21)	.49	4.3%
CORE Practice-Based Evidence National Database 2008[7]	UK	CORE-OM	1.78 (6.24)	0–3.85	0.87 (0.63)	0–3.64	Yes	6.83 (4.37)	1.45	10.2%

IAPT project[8]	UK	PHQ-9	14.78 (6.24)	1–27	9.15 (6.77)	1–27	No	6.63 (3.81)	.90	2.7%

Note. TK-project = Techniker Krankenkassen project; CORE = Clinical Outcomes in Routine Evaluation; IAPT = Improving Access to Psychological Therapies; GER = Germany; US = United States; UK = United Kingdom;; GER = Germany; US = United States; UK = United Kingdom; BSI = Brief Symptom Inventory; BHM = Behavior Health Measure; MHI = Mental Health Index; OQ = Outcome Questionnaire; CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure; PHQ-9 = Patient Health Questionnaire; TE = Therapist effect; [1]N = 668; [2]N = 636; [3]N = 752; [4]N = 11,356; [5]N = 1,194; [6]N = 2,561; [7]N = 25,842; [8]N = 5,639; [9]Number of sessions of German datasets were corrected for probatorical sessions.[10]Effect size = Cohen's d; [11]All presented TEs are baseline adjusted estimates.

Table 2

Three-level MLM – basic model controlled for initial impairment.

Parameter	Null model	Model 1
Fixed effects		
Intercept	-0.97***	-0.89***
Initial impairment		-0.50***
Random effects	Variance (SD)	Variance (SD)
Level 3	0.09 (0.30)	0.04 (0.20)
Level 2		
Therapist	0.04 (0.21)	0.05 (0.21)
Initial Impairment		0.02 (0.13)
Level 1	0.78 (0.88)	0.58 (0.76)

Note. Number of patients N_{Pat} = 48,648; Number of therapists N_{Ther} = 1,800; Number of datasets N_d = 8.

$^{***}p$ =.001 $^{**}p$ < .01. $^{*}p$ < .05. ^{+}p < .1.

Table 3

Sample size table

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
30	400	4.36	0.50	4.23	4.73
	200	4.32	1.46	3.93	5.39
	100	4.34	2.47	3.73	6.20
	50	4.33	3.91	3. 37	7.28
	30	4.41	5.39	3.03	8.42
	20	4.54	7.01	2.89	9.90
	10	4.96	11.86	2.25	14.11
	5	5.85	17.23	1.94	19.17
	2	8.53	26.98	2.05	29.03
25	400	4.22	0.59	4.11	4.70
	200	4.27	1.57	3.90	5.47
	100	4.33	2.40	3.69	6.09
	50	4.34	3.84	3.44	7.28
	30	4.52	5.32	3.19	8.51
	20	4.74	7.44	2.95	10.39
	10	5.15	11.89	2.22	14.11
	5	5.63	17.49	1.75	19.24
	2	8.44	11.72	17.99	29.71
20	400	4.50	0.76	4.28	5.04
	200	4.50	1.93	4.04	5.97
	100	4.61	2.86	3.93	6.79
	50	4.75	4.59	3.56	8.15
	30	4.79	6.28	3.25	9.53
	20	4.87	8.40	3.01	11.41
	10	5.66	13.68	2.66	16.34
	5	6.48	16.88	2.24	19.12
	2	8.89	33.88	2.3	36.18

(continued)

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
15	400	4.31	0.77	4.13	4.90
	200	4.39	1.93	3.88	5.81
	100	4.51	3.12	3.73	6.85
	50	4.68	4.96	3.53	8.49
	30	4.82	6.51	3.06	9.57
	20	5.04	8.38	3.12	11.5
	10	5.41	12.27	2.32	14.59
	5	6.67	19.39	2.12	21.51
	2	9.23	29.95	2.21	32.16
10	400	5.13	0.94	4.90	5.84
	200	5.15	2.55	4.54	7.09
	100	5.22	4.45	4.15	8.60
	50	5.38	6.73	3.74	10.47
	30	5.69	8.75	3.42	12.17
	20	6.00	11.99	3.16	15.15
	10	6.60	19.22	2.41	21.63
	5	8.52	26.82	2.46	29.28
	2	14.18	49.48	2.79	52.27
5	400	4.62	1.58	4.24	5.82
	200	4.64	4.06	3.61	7.67
	100	4.88	7.29	3.10	10.39
	50	5.49	11.35	2.90	14.25
	30	6.23	14.85	2.40	17.25
	20	6.91	17.3	2.52	19.82
	10	9.36	28.64	2.37	31.01
	5	11.86	37.65	3.11	40.76
	2	20.56	65.46	3.73	69.19

(continued)

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
4	400	5.43	2.10	4.89	6.99
	200	5.47	5.60	4.02	9.62
	100	5.64	8.71	3.31	12.02
	50	6.17	12.65	2.88	15.53
	30	6.79	18.87	2.21	21.08
	20	8.28	25.7	2.46	28.16
	10	11.44	35.75	3.36	39.11
	5	14.62	54.23	2.48	56.71
	2	26.35	72.97	6.44	79.41
3	400	7.18	2.55	6.56	9.11
	200	7.35	6.47	5.80	12.27
	100	8.10	13.20	5.10	18.3
	50	8.50	18.72	4.19	22.91
	30	9.99	23.98	4.41	28.39
	20	11.77	32.98	3.95	36.93
	10	14.95	46.5	3.77	50.27
	5	20.05	61.18	4.07	65.25
	2	34.74	84.35	8.77	93.12

Note. TE = therapist effect; TE is the mean therapist effect of 1,000 samples. Due to shortage of space and clarity not all sample size conditions are presented in the table.

Table 4

Multiple regression analysis with therapist effect (estimated per sample size condition for 1,000 samples) as outcome variable

Variable	*B (SE)*	95% *CI*
Constant	13.99*** (0.70)	[12.62, 15.37]
Patients per therapist	-0.36*** (0.04)	[-0.44, -0.27]
Therapists per dataset	-0.03*** (0.01)	[-0.04, -0.02]
Patients per therapist * therapist per dataset	0.001*** (0.00)	[0.00, 0.002]
R^2	0.44	
F-statistic	33.90***	

Note. SE = standard error; CI = confidence interval.

***$p < .001$

Appendix

Two-level hierarchical model

Level 1 (Patient Level): $\text{Outcome}_{\text{post ij}} = \pi_{0j} + \pi_{1j} * \text{initial impairment\_ centered}_{ij} + e_{ij}$

Level 2 (Therapist Level): $\pi_{0j} = \beta_{00} + r_{0j}$

$$\pi_{1j} = \beta_{10} + r_{1j}$$

Three-level hierarchical model

Level 1 (Patient Level): $\text{Outcome}_{\text{post ijk}} = \pi_{0jk} + \pi_{1jk} * \text{initial impairment\_ centered}_{ijk} + e_{ijk}$

Level 2 (Therapist Level): $\pi_{0jk} = \beta_{00k} + r_{0jk}$

$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$

Level 3 (Dataset Level): $\beta_{00k} = \gamma_{000} + u_{00k}$

$$\beta_{10k} = \gamma_{100} + u_{10k}$$

Note. MLM formulas for the hierarchical models predicting treatment outcome where patient *i* is nested within therapist *j* and therapist *j* is nested within dataset *k*. For each of the eight datasets, initial impairment was standardized on the mean and standard deviation of an appropriate country-specific outpatient reference sample (initial impairment_centered; see footnote 1) and included as a predictor on level 1 in order to capture the individual patient's psychological distress at intake as a deviation from the relevant population mean. Considering the AIC a random intercept (r_{0jk}; u_{00k}) and random slope (r_{1jk}; u_{10k}) model consistently fit the data best.