Qual Quant (2012) 46:1699–1708
DOI 10.1007/s11135-011-9548-4

Benchmarking therapists: furthering the benchmarking method in its application to clinical practice

Takuya Minami · G.S Brown · Joyce McCulloch · Brent J. Bolstrom

Published online: 7 September 2011

Abstract Psychotherapy research has been interested in understanding the variability observed among therapists with regard to their treatment effectiveness. An important initial step towards understanding the source of the differences is to reliably identify therapists that are effective. The current paper thus proposes a method for benchmarking therapists against predetermined criteria of effectiveness which could be conducted using any standard statistical package. Basic steps include (a) creating benchmark(s), (b) determining a prior the numerical criteria that constitute as "effective" based on the benchmark(s), (c) calculating pre-post effect sizes as an indicator of effectiveness at the case level using statistical adjustments so as to best match clinical (and other) differences among cases, and (d) statistically benchmarking the therapists using a random-effects hierarchical linear modeling. An example is provided that highlights the number of therapists who would be classified as effective based on various numerical criteria and confidence levels.

Keywords Bench marking · Therapist effects · Effectiveness · Psychotherapy · Clinical practice

1 Introduction

Assessing outcomes of psychotherapy practiced in natural clinical settings (psychotherapy treatment-as-usual; PTAU) has been considered one of the most important questions in

T. Minami (✉)
Department of Counseling and Applied Educational Psychology, Northeastern University,
Boston, MA, USA
e-mail: t.minami@neu.edu

G. S. Brown
Center for Clinical Informatics, Salt Lake City, UT, USA

J. McCulloch · B. J. Bolstrom
United Behavioral Health, San Francisco, CA, USA

psychotherapy research since almost half a century ago (e.g., Cohen 1965; Luborsky 1972; Seligman 1995; Strupp 1989). However, very few studies have investigated PTAU effectiveness, and when conducted, divergent research methods and their limitations have defied systematic accumulation of evidence (Minami and Wampold 2008). For example, clinical representativeness studies, although innovative, cannot be considered solid evidence because PTAU effectiveness is statistically estimated using data from clinical trials rather than directly measuring treatment outcomes (e.g., Shadish et al. 1997, 2000; Shapiro and Shapiro 1982). Other studies have assessed PTAU effectiveness by comparing PTAU against empirically supported treatments (ESTs) that were transported into natural clinical settings (e.g., Addis et al. 2004; Linehan et al. 1999; Rawson et al. 2004). It is difficult to draw conclusions from these studies because of multiple issues including unequal number of sessions, differences in amount of training and supervision, and higher expectations of effectiveness for ESTs over PTAUs.

One analytical method, namely benchmarking, appears to be a promising method for assessing PTAU effectiveness. The methodology for benchmarking PTAU outcomes has been developed over the past decade (Merrill et al. 2003; Minami et al. 2008a; Wade et al. 1998; Weersing and Weisz 2002). One recently published method for benchmarking involves three steps: (a) meta-analytically aggregating clinical trials to construct a benchmark, (b) measuring PTAU effectiveness, and (c) statistically comparing the PTAU effectiveness to the benchmark (Minami et al. 2008a). Using this method, PTAU effectiveness has thus far been primarily assessed in a managed care environment (Minami et al. 2008b), university counseling center (Minami et al. 2009), and community-based treatment for juvenile offenders (Curtis et al. 2009).

However, in addition to the above benchmarking method being appropriate only for large samples (Minami et al. 2007), it does not take into account the possible differences among individual therapists in their effectiveness. Although this is not an issue if the purpose of the benchmarking is to evaluate PTAU effectiveness at the level of organizations (e.g., a community mental health center), the method is likely impractical for assessing effectiveness at the level of the therapist. Evaluation of effectiveness at the level of the therapist has been of significant interest for psychotherapy researchers for decades (e.g., Durlak 1979; Luborsky 1952; Martindale 1978; Rosenzweig 1936). Recently, with the advancement of statistical methods, evidence of variability among therapists in their treatment outcomes has been robust both in natural clinical settings and in controlled trials (e.g., Huppert et al. 2001; Kim et al. 2006; Okiishi et al. 2003; Wampold and Brown 2005). Therefore, it is crucial that PTAU effectiveness is assessed at the level of the therapist using credible benchmarks.

2 Benchmarking therapists

In most psychotherapy data, treatment data obtained from clients are not independent from one another. This is because the same therapist sees more than one client, and therefore, there often is variance that is due to the therapist that is independent of the client. Such data structure is known as nesting because observations at one level (e.g., client) are contained within a particular observation at another level (e.g., therapist).

When assessing PTAU effectiveness among therapists, the issues surrounding the nested data structure become more pronounced because therapists may indeed see different types of clients. For example, therapists may specialize based on type of issues (e.g., major depression, eating disorders, substance abuse) or population (e.g., with regards to age, socioeconomic status). This requires benchmarking PTAU effectiveness at the level of therapists to differ from benchmarking at the organizational level because equating the treatment outcomes across therapists becomes a crucial issue. Therefore, therapist benchmarking consists of four major steps: (a) constructing the benchmarks, (b) determining what effect size could be considered "effective," (c) calculating PTAU effectiveness at the case level, and (d) statistically benchmarking the therapists based on the predetermined effect size criteria of effectiveness. In particular, the last two steps are crucial modifications to the current benchmarking method so as to accommodate the nested structure of the data.

2.1 Constructing benchmarks

As illustrated in detail in Minami et al. (2008a), currently the best benchmarks are meta-analytic aggregations of pre-post treatment outcome in well-conducted clinical trials. Several key issues in constructing the benchmarks are (a) predetermined inclusion/exclusion and coding criteria, (b) type of posttest (i.e., "completers" or "intent-to-treat"), and (c) type of outcome assessment (i.e., reactivity and specificity).

It is important that clear inclusion/exclusion criteria are determined a priori so that researchers need not contemplate which studies should be included for their benchmarks while searching the literature. It is also important that possible moderating variables are identified prior to reviewing the individual study so as to quickly identify and code them appropriately. These factors are crucial in benchmarking because it is more often than not that the clinical characteristics of the clients in clinical trials are different from the clients receiving PTAU (see also Sect. 2.3 below). For example, if health status is available for the clients receiving PTAU, this is also a variable that should be coded when constructing the benchmark so that the benchmark and the PTAU data could be matched as best possible.

As evident from Minami et al. (2007) in their construction of benchmarks for adult depression treatment, pre-post treatment effect sizes differ depending on whether the clients in the clinical trials completed the full duration of treatment or terminated early. Given that most PTAU data do not contain information on the type of treatment, it is crucial that the benchmarks are constructed with clinical trials that have treatment outcomes on intent-to-treat population than just completers. Another factor that affects pre-post treatment effect size is the type of outcome measure with regard to (a) reactivity (i.e., clinician-administered or client self-report) and (b) specificity (i.e., tailored to a specific diagnosis or not; Smith et al. 1980; Lambert et al. 1986). It is ideal if there are enough clinical trials with identical outcome measures to construct a benchmark; however, if not, then the outcome measures should at least be matched based on reactivity and specificity.

2.2 Determining what effect size could be considered as "effective"

Once the benchmark(s) are constructed, the researcher is confronted with a difficult decision—determining what magnitude of effect size is sufficient to be considered effective (or any other designation). For example, given an intent-to-treat treatment efficacy benchmark of $d_{B(TE)} = 0.80$ based on an aggregate of low reactivity and low specificity outcome measures (Minami et al. 2007), would an observed treatment effect size of $d = 0.75$ be considered large enough to consider the therapist as effective? How about $d = 0.65$? Potential remedies for this issue are reflected in recent benchmarking studies. For example, in Curtis et al. 2009 and Minami et al. (2008a), a margin of $d_{\Delta} = 0.2$ (i.e., below the clinical trials benchmark) was considered as "good enough" (Serlin and Lapsley 1985, 1993) by following Cohen (1988) recommendation of a small effect size. On the other hand, Minami et al. (2008b) and

Minami et al. (2009) chose a margin of 10% of the efficacy benchmark (i.e., $d_{\Delta} = 0.08$ below the benchmark) so as to be more conservative with their analysis.

However, deciding what constitutes effectiveness that is “good enough” cannot be determined solely on statistical rationale. What is necessary is to consider the practical consequences of using a certain numerical margin as a clinically meaningful criterion in light of the benchmark. For example, Minami et al. (2007) reported a benchmark for natural remission benchmark of $d_{B(NR)} = 0.15$ based on low reactivity and low specificity outcome measures. In other words, this is the effect size observed in clinical trials if clients were to be randomized to the control group. Given also the $d_{B(TE)} = 0.80$ for the treatment efficacy benchmark, if the observed PTAU produced an effect size of $d = 0.30$, should this treatment be considered clinically effective? What if the observed PTAU effect size was $d = 0.50$ or $d = 0.70$? The issue here cannot be resolved statistically because with a sufficient number of clients in the database, any margin of effect size would result in statistical significance. Therefore, a rather subjective decision must be made not only based on the calculated benchmarks but also upon the intended use of the effectiveness assessment. For example, a PTAU effect size of $d = 0.30$, while seemingly small, may still be regarded as useful if the treatment was a single-session telephonic intervention for clients in remote areas who cannot attend regular therapy sessions. In addition, the criteria for effectiveness need not be one value. For example, one could designate an effect size under $d = 0.30$ as “not very effective,” values between $d = 0.30$ and 0.70 as “effective,” and effect sizes above $d = 0.70$ as “very effective,” which then leads to a 3-tiered benchmarking. Again, these criteria need to be determined prior to analyses and based on the purpose of assessment.

2.3 Calculating PTAU effectiveness at the case level

As illustrated in Minami et al. (2008a), benchmarking PTAU effectiveness at the organization level involves calculating the PTAU effect size as a whole rather than by case. In doing so, a major issue for natural clinical settings is that the clients’ clinical (and other) characteristics often do not match that of the clients included in the benchmarks, especially if the benchmarks were obtained from clinical trials. In many clinical trials, it is crucial that strict inclusion/exclusion criteria result in participants with homogeneous clinical characteristics. This implies, however, that when the assessment of PTAU effectiveness is conducted on a client population with heterogeneous clinical characteristics, the difference in clinical characteristics between the benchmark and the PTAU needs to be taken into consideration.

Because most clinical trials that would qualify for inclusion in constructing a benchmark are based on clients with rather homogeneous characteristics, it is ideal that PTAU data include variables that may potentially impact treatment outcomes, such as diagnosis, socioeconomic status, and physical health. If these data are available, the impact of such factors could be statistically assessed and controlled. For example, it is well documented in the clinical literature that the initial level of distress significantly predicts pre-post treatment outcome (e.g., Clarkin and Levy 2005; Lambert 2001). If the clinical trials benchmark and the PTAU data utilize identical outcome measures, PTAU data could be easily adjusted to match the severity of distress observed in the clinical trials using multiple regression prior to benchmarking the data. Other factors (e.g., diagnosis) could also be statistically controlled in the same manner.

When benchmarking at the therapist level, it becomes increasingly important that statistical adjustments be made with the PTAU data even in the case that the outcome measures are not identical to the benchmark. This is because it is very likely that different therapists see clients with idiosyncratic clinical characteristics—for example, some therapists may specialize in anxiety while another may focus on eating disorders. In such a case, it may be grossly misleading to compare one therapist's caseload to another because of the differences in clinical characteristics. If the PTAU database as a whole is large enough, reliable statistical adjustments (often referred to as case-mix adjustments) are possible so as to take into consideration these differences in clinical and other characteristics even if each therapist's caseload on average may be small. In this instance, statistical adjustment of each case (rather than the overall effect size of the PTAU data) based on clinical characteristics becomes possible so that the benchmarking is done using severity adjusted effect sizes (SAES) rather than raw (observed) effect sizes. Therefore, when statistically benchmarking therapists, effect sizes need to be calculated at the case level rather than the dataset as a whole.

SAES for each PTAU case is calculated in three steps: (a) conduct a multiple regression with the raw effect size as the dependent variable and clinical characteristics (and other potentially moderating) factors as independent variables, (b) save the residual for each case (i.e., raw effect size minus the predicted effect size), and (c) add the overall intercept and the residual to derive the SAES. What this process does is that every case is now statistically adjusted for all moderators and thus could be compared against one another.

2.4 Statistically benchmarking therapists

Once the benchmark(s) are constructed, decisions have been made as to the effect sizes that would be considered as effective, and the SAESs are calculated for each PTAU case, the last step is to estimate each therapists' mean SAES to compare against the criteria of effectiveness. To reflect the nested nature of the data, hierarchical linear modeling (HLM; e.g., Raudenbush and Bryk 2002) should be used. HLM could be conducted using any common statistical packages such as SAS, SPSS, HLM, Stata, S-PLUS, and R.

An additional issue here is whether to consider the therapists fixed or random. From a conceptual standpoint, it is defensible to consider therapists as a fixed factor only if the interest is to benchmark only these therapists given the provided set of data. However, two main reasons suggest that therapists should be considered a random factor in most cases. First, if there is any interest in generalizing the results, it would be erroneous to consider the therapists as fixed (Martindale 1978; Wampold and Serlin 2000). In particular, Serlin et al. (2003) note that "[c]onclusions drawn on the basis of [treating therapists as a random factor] can be generalized …to providers not included in the study, and to subsequent administrations of the treatment by the same providers at other points in time or in different situations (e.g., in a different office)" (p. 528). Second issue is a rather pragmatic one; in a fixed-factor HLM, estimates of therapists' mean effect sizes are simple unweighted averages of their clients' SAESs. This obviously is very problematic in situations where therapists do not have many cases in the database (e.g., less than 10) because the significant variability among clients directly affects the therapists' average SAES (e.g., some therapists had a bad streak). This unreliability can be disturbing for those therapists whose simple average SAESs are far from the overall average. For example, in the case of a 3-tier benchmarking of effectiveness, the consequences for a therapist to be classified as "not very effective" when indeed he/she might be effective in reality can be highly problematic. Similarly, it could be disturbing for therapists to be rated "highly effective" only to be reclassified as "effective" after a few additional clients. The pragmatic benefit of the random-effects model is that, as compared to a fixed-effects model, the therapists' mean SAESs are estimated conservatively (i.e., closer to the mean of all cases) for those therapists with lower number of cases and farther away from the mean. Therefore, therapists' mean SAESs estimated using a random-effects model inherently creates a "benefit-of-the-doubt" effect to those therapists who have a lower simple mean SAES.

Using a random-effects HLM, therapists' mean SAES are estimated along with the standard error of the estimate. Rather than using solely the estimated mean SAES, a confidence interval is created around the mean using the standard error. Whereas social sciences research invariably utilizes a confidence level of 95%, the confidence level could also be adjusted depending on the purpose of the benchmarking. For example, a 95% confidence may not be necessary if the purpose is to identify therapists who are likely to have decent average treatment effect. On the other hand, if the purpose is to identify therapists who, on average, are quite ineffective, then a high level of confidence would be a necessity given the potential implications as well as an effectiveness criterion that is justifiable. In addition, the confidence interval need not be two-sided if only one end of the confidence limit is of interest. For example, if the interest is only in identifying therapists that exceed a certain minimum threshold, only the lower-bound confidence limit is necessary.

3 Illustration

As an example, we provide here a simulation of the benchmarking method using data obtained from clients seen in a managed care environment. The database contained treatment episode data for 54,753 clients seen by 17,152 therapists from December 2006 to July 2009. The number of clients under each therapists was positively skewed, with an overall mean of $n = 3.19$, median of 2, mode of 1, and range of 1 to 90. Because the database was authorized solely for the purpose of this simulation, the database was deidentified, containing no other clinical or demographic information on the clients or therapists. However, statistical adjustments had been performed at the case level prior to de-identification. General linear models were used to predict the expected change score for each case. The model included a number of independent variables determined to be predictive of change including, but not limited to, baseline severity, time between measurements, and patient characteristics including age and gender. Although the number of sessions is known to be predictive of the overall pre-post treatment effect size, it was intentionally left uncontrolled as the interest was on clients' clinical improvement regardless of how many sessions it may have taken (e.g., Baldwin et al. 2009). SAES was computed for each treatment case and included in the de-identified dataset. For this example, all statistical adjustments and analyses were conducted using SAS.

Therapist benchmarking is illustrated here using a single cutoff criterion, varied from $d_C = 0.15$ (i.e., the natural remission benchmark) to $d_C = 0.80$ (i.e., the treatment efficacy benchmark) as the indicator of effectiveness. The confidence level was also varied from 70% to 99.5% (one-tailed). Thus, if a therapist's lower-bound confidence level exceeds the pre-specified d_C, then the therapist is designated as "effective." If the therapist's lower-bound confidence level does not exceed this level, the therapist is simply not given any designation because the reasons for not exceeding this threshold may be several, including not having enough cases.

Designating 10 cases as the minimum caseload for therapists resulted in including 1,003 (out of a total of 17,152) therapists in the benchmarking. Therapists' mean SAES were approximately normally distributed (Fig. 1), ranging from $d_{SAES} = 0.457$ to $d_{SAES} = 1.222$. Their 95% lower-bound confidence limits ranged from $d_L = 0.235$ to $d_L = 1.001$. Table 1 illustrates the number of therapists that are designated as "effective" given a certain criterion (d_C) and confidence level (%). For example, when $d_C = 0.50$ at 95% confidence level results in 659 (65.70%) therapists being classified as "effective." Note, for example, that if the

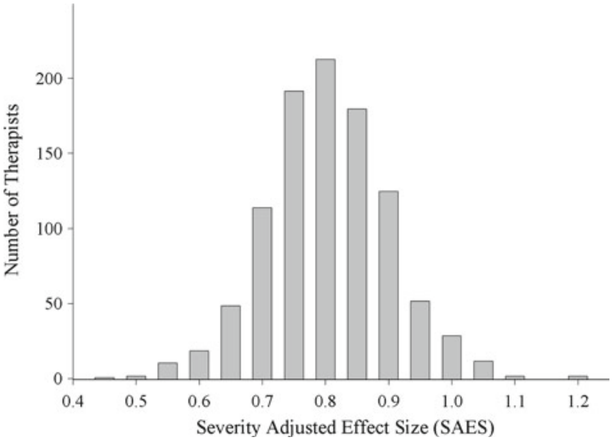

Fig. 1 Distribution of therapists' mean SAES ($N = 1,003$)

Table 1 Number (of $N = 1,003$) of therapists designated as effective as function of effectiveness criteria and confidence level ($n = 10+$)

Criteria (d_C)	Confidence level								
	70.0%	75.0%	80.0%	85.0%	90.0%	95.0%	97.5%	99.0%	99.5%
0.15	1003	1003	1003	1003	1003	1003	1003	997	988
0.20	1003	1003	1003	1003	1003	1003	1001	992	947
0.25	1003	1003	1003	1003	1003	1001	995	954	856
0.30	1003	1003	1003	1003	1002	995	977	873	691
0.35	1003	1003	1002	1001	997	983	925	716	500
0.40	1002	1002	1000	995	985	937	814	523	310
0.45	999	997	991	982	952	835	627	332	180
0.50	990	982	971	940	877	659	424	172	88
0.55	970	948	917	851	723	452	227	85	41
0.60	916	871	800	677	504	256	114	44	22
0.65	793	712	598	463	306	124	53	20	6
0.70	594	493	394	261	143	55	25	6	4
0.75	388	295	208	128	67	28	7	4	2
0.80	205	143	97	56	33	7	4	2	2

As the total number of therapists included in the benchmarking is $N = 1,003$, the number of therapists divided by 10 in each cell is roughly the percentage of therapists in that cell

confidence level was at 90%, the same cutoff score would result in 877 (87.44%) therapists being classified as "effective." In most practical applications, it is very unlikely that a 95% confidence, which is often used in research, is necessary. Rather, a 90% confidence may be a feasible upper limit of confidence.

4 Summary and Conclusion

In the past decade, there has been a renewed interest in investigating the variability among therapists with regards to their clinical outcomes. Our current study thus attempted to further the benchmarking method beyond the aggregate level (as in Minami et al. 2008a) to the therapist level. Main improvements include (a) individually adjusting the clients' cases by calculating the SAES so as to statistically best match the benchmark prior to conducting the analysis, (b) clarifying the process to determine numerical criteria for effectiveness, and (c) utilizing random-effects HLM that has been available in many standard statistical packages for some time.

As with any group comparisons, caution is necessary when applying the benchmarking method. First, because the comparisons are different from that of between-group comparisons among different conditions in a randomized clinical trial, statistically controlling for differences in clinical (and other) characteristics in the PTAU data does not assure that the characteristics between the clients in the benchmarks and the PTAU data are similar enough. Therefore, depending on the nature of the PTAU data, it is crucial that benchmarks are created after a thorough review of the available evidence. Second, the statistical control also cannot take into consideration every single difference among clients in the PTAU data, especially if such data are not collected. Therefore, although using SAES is justified over using raw effect sizes to compare client cases against one another, there nevertheless may be other systematic differences among client groups that may not be taken into account. Although using a random-effects model provides a "benefit-of-the-doubt" effect for therapists with mean estimated SAES that is lower than the overall average, it is still necessary to be cautionary with regards to interpreting the benchmarking results. An inherent issue with any statistical analysis is that all estimations are based on the sample at hand rather than the population (which can never be known unless in very unusual cases); therefore, the degree to which a particular therapists' mean estimated SAES is adjusted towards the overall mean is partly a function of the rest of the therapists and clients in the dataset. Third, it is important to note again that therapist SAESs not surpassing a lower-bound (or upper-bound) confidence limit indicates that there is insufficient evidence to consider the estimated effectiveness as surpassing this limit. This interpretation is clearly different than to state that there is evidence that the therapist does not meet the criterion of effectiveness.

For both researchers and organizations, there are compelling reasons as to why benchmarking therapists may be beneficial. For researchers, a crucial question is to identify why some therapists excel in providing psychotherapy. What we might learn from these investigations could significantly impact the nature of our clinical training and continuing education efforts. For organizations, it may be to their advantage to identify therapists who may not be performing at the desired level so that the organization could provide further support (e.g., increase number of sessions for cases with low outcome; Baldwin et al. 2009). Understanding why some therapists are effective would be a crucial step in increasing the benefit of psychotherapy to our clients.

References

Addis, M.E., Hatgis, C., Krasnow, A.D., Jacob, K., Bourne, L., Mansfield, A.: Effectiveness of cognitive-behavioral treatment for panic disorder versus treatment as usual in a managed care setting. J. Consult. Clin. Psychol. **72**, 625–635 (2004)

Baldwin, S.A., Berkeljon, A., Atkins, D.C., Olsen, J.A., Nielsen, S.L.: Rates of change in naturalistic psychotherapy: contrasting dose-effect and good-enough level models of change. J. Consult. Clin. Psychol. **77**, 203–211 (2009)

Clarkin, J.F., Levy, K.N.: The influence of client variables on psychotherapy. In: Lambert, M.J. (ed.) Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, Wiley, New York (2005)

Cohen, J.Some statistical issues in psychological research. In: Wolman, B. (ed.) Handbook of Clinical Psychology, McGraw-Hill, New York (1965)

Cohen, J.: Statistical Power Analysis for the Behavioral Sciences.95–121 2nd edn. Lawrence Erlbaum Associates, Hillsdale (1988)

Curtis, N.M., Ronan, K.R., Heiblum, N., Crellin, K.: Dissemination and effectiveness of multisystemic treatment in New Zealand: a benchmarking study. J. Fam. Psychol. **23**, 119–129 (2009)

Durlak, J.A.: Comparative effectiveness of paraprofessional and professional helpers. Psychol. Bull. **86**, 80–92 (1979)

Huppert, J.D., Bufka, L.F., Barlow, D.H., Gorman, J.M., Shear, M.K., Woods, S.W.: Therapists, therapist variables, and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. J. Consult. Clin. Psychol. **69**, 747–755 (2001)

Kim, D.M., Wampold, B.E., Bolt, D.M.: Therapist effects in psychotherapy: a random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. Psychother. Res. **16**, 161–172 (2006)

Lambert, M.J.: The status of empirically supported therapies: comment on Westen and Morrison's (2001) multidimensional meta-analysis. J. Consult. Clin. Psychol. **69**, 910–913 (2001)

Lambert, M.J., Hatch, D.R., Kingston, M.D., Edwards, B.C.: Zung, Beck, and Hamilton rating scales as measures of treatment outcome: a meta-analytic comparison. J. Consult. Clin. Psychol. **54**, 54–59 (1986)

Linehan, M.M., Schmidt, H., Dimeff, L.A., Craft, J.C., Kanter, J., Comtois, K.A.: Dialectical behavior therapy for patients with borderline personality disorder and drug-dependence. Am. J. Addict. **8**, 279–292 (1999)

Luborsky, L.: The personality of the psychotherapist. Menninger Q. **6**, 1–6 (1952)

Luborsky, L.: Research cannot yet influence clinical practice. In: Bergin, A., Strupp, H. (eds.) Changing Frontiers in Science of Psychotherapy, pp. 120–127. Aldine, Chicago (1972)

Martindale, C.: The therapist-as-fixed-effect fallacy in psychotherapy research. J. Consult. Clin. Psychol. **46**, 1526–1530 (1978)

Merrill, K.A., Tolbert, V.E., Wade, W.A.: Effectiveness of cognitive therapy for depression in a community mental health center: a benchmarking study. J. Consult. Clin. Psychol. **71**, 404–409 (2003)

Minami, T., Wampold, B.E.: Adult psychotherapy in the real world. In: Walsh, W.B. (ed.) Biennial Review of Counseling Psychology, pp. 27–45. Routledge, New York (2008)

Minami, T., Wampold, B.E., Serlin, R.C., Kircher, J.C., Brown, G.S.: Benchmarks for psychotherapy efficacy in adult major depression. J. Consult. Clin. Psychol. **75**, 232–243 (2007)

Minami, T., Serlin, R.C., Wampold, B.E., Kircher, J.C., Brown, G.S.: Using clinical trials to benchmark effects produced in clinical practice. Qual. Quant. **42**, 513–525 (2008a)

Minami, T., Wampold, B.E., Serlin, R.C., Hamilton, E.G., Brown, G.S., Kircher, J.C.: Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: a preliminary study. J. Consult. Clin. Psychol. **76**, 116–124 (2008b)

Minami, T., Davies, D.R., Tierney, S.C., Bettmann, J.E., McAward, S.M., Averill, L.A., Huebner, L.A., Weitzman, L.M., Benbrook, A.R., Serlin, R.C., Wampold, B.E.: Preliminary evidence on the effectiveness of psychological treatments delivered at a university counseling center. J. Couns. Psychol. **56**, 309–320 (2009)

Okiishi, J., Lambert, M.J., Nielsen, S.L., Ogles, B.M.: Waiting for supershrink: an empirical analysis of therapist effects. Clin. Psychol. Psychother. **10**, 361–373 (2003)

Raudenbush, S.W., Bryk, A.S.: Hierarchical Linear Models. 2nd edn. Sage, Newbury Park (2002)

Rawson, R.A., Marinelli-Casey, P., Anglin, M.D., Dickow, A., Frazier, Y., Gallagher, C., Galloway, G.P., Herrell, J., Huber, A., McCann, M.J., Obert, J., Pennell, S., Reiber, C., Vandersloot, D., Zweben, J., Methamphetamine Treatment Project Corporate Authors: A multi-site comparison of psychosocial approaches for the treatment of methamphetamine dependence. Addiction **99**, 708–717 (2004)

Rosenzweig, S.: Some implicit common factors in diverse methods of psychotherapy. Am. J. Orthopsychiatr. **6**, 412–415 (1936)

Seligman, M.E.P.: The effectiveness of psychotherapy: the consumer reports study. Am. Psychol. **50**, 965–974 (1995)

Serlin, R.C., Lapsley, D.K.: Rationality in psychological research: the good-enough principle. Am. Psychol. **40**, 73–83 (1985)

Serlin, R.C., Lapsley, D.K.: Rational appraisal of psychological research and the good-enough principle. In: Keren, G., Lewis, C. (eds.) A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues, pp. 199–228. Lawrence Erlbaum Associates, Hillsdale (1993)

Serlin, R.C., Wampold, B.E., Levin, J.R.: Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: a comment on Siemer and Joormann (2003). Psychol. Methods **8**, 524–534 (2003)

Shadish, W.R., Matt, G.E., Navarro, A.M., Siegle, G., Crits-Christoph, P., Hazelrigg, M.D., Jorm, A.F., Lyons, L.C., Nietzel, M.T., Prout, H.T., Robinson, L., Smith, M.L., Svartberg, M., Weiss, B.: Evidence that therapy works in clinically representative conditions. J. Consult. Clin. Psychol. **65**, 355–365 (1997)

Shadish, W.R., Matt, G.E., Navarro, A.M., Phillips, G.: The effects of psychological therapies under clinically representative conditions: a meta-analysis. Psychol. Bull. **126**, 512–529 (2000)

Shapiro, D.A., Shapiro, D.: Meta-analysis of comparative therapy outcome studies: a replication and refinement. Psychol. Bull. **92**, 581–604 (1982)

Smith, M.L., Glass, G.V., Miller, T.I.: The Benefits of Psychotherapy. Johns Hopkins University Press, Baltimore (1980)

Strupp, H.H.: Psychotherapy: can the practitioner learn from the researcher? Am. Psychol. **44**, 717–724 (1989)

Wade, W.A., Treat, T.A., Stuart, G.L.: Transporting an empirically supported treatment for panic disorder to a service clinic setting: a benchmarking strategy. J. Consult. Clin. Psychol. **66**, 231–239 (1998)

Wampold, B.E., Serlin, R.C.: The consequences of ignoring a nested factor on measures of effect size in analysis of variance designs. Psychol. Methods **4**, 425–433 (2000)

Wampold, B.E., Brown, G.S.: Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. J. Consult. Clin. Psychol. **73**, 914–923 (2005)

Weersing, V.R., Weisz, J.R.: Community clinic treatment of depressed youth: benchmarking usual care against CBT clinical trials. J. Consult. Clin. Psychol. **70**, 299–310 (2002)