

## ACORN Criteria for Effectiveness (ACE) Treatment Outcomes Evaluation Method

Prepared by: Takuya Minami, PhD & Jeb Brown, PhD

### Purpose

This document illustrates a method of qualitatively interpreting psychotherapy outcomes as assessed by quantitative measures of clients' self-reported clinical symptoms that are administered repeatedly throughout the course of therapy. For any qualitative interpretations of quantitative data, it becomes necessary to categorize the continuous quantitative outcomes by using cutoff scores. For example, it becomes necessary to quantitatively determine what magnitude of change could be categorized as "improved." Or, it requires a criterion to designate a particular clinician as "effective" based on one's clients' average outcomes. Therefore, to qualitatively interpret any quantitative data, it is necessary that cutoff scores are utilized.

Although any cutoff score is arguably arbitrary to some extent, criteria derived based on knowledge of effect size and psychometrics are better defended. In addition, qualitative descriptions of treatment effect must ultimately be meaningful to both the clinician and the client. Therefore, any qualitative interpretations using cutoff scores must be intuitive yet psychometrically sound.

Of many possible methods to qualitatively interpret quantitative treatment outcome data, here are currently two methods that we find are worth recommending: caseload assessment using the standard error of measurement and benchmarking using standardized pre-post change score. To comprehend these two methods, the following five components require understanding: standard error of measurement (*SE*), clinical cutoff score, reliable change index (*RCI*), standardized pre-post mean change (Cohen's *d*), and residuals.

### Psychometric Background

#### A. Standard Error of Measurement (*SE*)

In any repeated measurement, random perturbations occur. Thus, while an assessment is a good estimate, it is likely different from the "true" score. *SE* is the estimate of (random) error inherent in the particular assessment that is taken into consideration when interpreting an individual's score. This is estimated as below:

$$SE = SD\sqrt{1 - r_{xx}} \quad (1)$$

Here, *SD* is simply the standard deviation of the measure, and  $r_{xx}$  is the reliability estimate of the measure (e.g., Cronbach's alpha). The individual's "true" score could be expected to fall within 1 *SE* above or below the observed score approximately 68% of the time, and within 2 *SE* above or below approximately 95% of the time.

*SE* is important in interpreting clients' progress because we are interested in "true" change occurring between assessments rather than random fluctuations due to measurement error.

Therefore, to claim any “true” change, the difference between two assessments must be beyond 1 *SE* to claim any possibility of “true” change in clinical symptoms.

### *B. Clinical Cutoff Score and Reliable Change Index (RCI)*

Jacobson and Truax (1991) proposed two statistical criteria that could be considered as indicators of change: clinical cutoff score and reliable change index (RCI).

The clinical cutoff score indicates a point on a clinical outcome measure that could be considered as a criterion between clients who could be considered to be within a clinical range and those in a nonclinical range. Specifically:

$$C = \frac{SD_1 \cdot mean_2 + SD_2 \cdot mean_1}{SD_1 + SD_2} \quad (2)$$

Here,  $mean_1$  ( $SD_1$ ) and  $mean_2$  ( $SD_2$ ) are the means and standard deviations of the clinical and nonclinical samples, respectively. Jacobson and Truax proposed that when a client’s score crosses this cutoff score, it is likely that clients have attained reliable change (hopefully in the direction of clinical to nonclinical).

The clinical cutoff score is important in any evaluation of improvement because patients who come into treatment in the normal range should be excluded for assessment as most commonly used outcome measures are not sensitive enough to detect changes within the nonclinical range. In addition, limiting treatment outcome evaluations to those whose level of distress is within the clinical range would more resemble patients who are included in clinical trials, thus allowing for evaluation against clinical trials (see also Benchmarking). In a typical outpatient treatment population, 20-30% of patients present with test scores in the nonclinical range at the beginning of treatment.

A simple crossing of the clinical cutoff score may not be sufficient to claim reliable change as (a) the clinical and nonclinical samples may have considerable overlap and (b) while crossing the cutoff score, small changes may indeed be indicative of random fluctuations. In attempt to answer what magnitude of change could reasonably be classified as reliable, Jacobson and Truax originally proposed the reliable change index ( $RCI_{orig}$ ) as below:

$$RCI_{orig} = \frac{x_1 - x_2}{\sqrt{2 \cdot SE^2}} \quad (3)$$

Here,  $x_1$  and  $x_2$  are the client’s raw scores, and  $SE$  is the standard error of measurement explained above. When  $RCI_{orig}$  is greater than 1.96, it is unlikely, with a 95% confidence, that the magnitude of between the two points is simply due to random perturbation and thus could be interpreted as true change.  $RCI_{orig}$  was proposed so that the obtained value is scale independent (i.e., the value is not dependent on the scaling of the original measure). However, for purposes of use in actual clinical settings, it is more informative to know what actual difference in scores between two measurements would likely constitute “true” change with a 95% confidence. As the difference between  $x_1$  and  $x_2$  needs to be at or larger than  $RCI_{orig} = 1.96$  to claim reliable change, the required magnitude of difference between  $x_1$  and  $x_2$  (RCI) can be determined as below:

$$1.96 = \frac{RCI}{\sqrt{2 \cdot SE^2}}$$

$$\therefore RCI = 1.96\sqrt{2 \cdot SE^2} \quad (4)$$

Here, RCI would be in the units of the scale in use. For example, if the assessment was the ACORN Brief Adult Outcome Questionnaire (BAOQ), then the RCI would be the actual points on the BAOQ between two assessments that would likely, again with a 95% confidence, constitute “true” change. Obviously, as the RCI is a function of the *SE*, their ratio is a constant (i.e.,  $RCI/SE = 1.96 \cdot \sqrt{2}$ ).

There are two practical disadvantages in solely utilizing the RCI criteria for improvement. First, due to regression to the mean, the probability that a patient would improve (i.e., surpass the RCI in change) is higher for those whose severity of distress is higher at intake. Second, the 95% probability criterion for determining “true” change is rather conservative, and thus a large percentage of patients would likely be classified as “unchanged” simply because of the strict Type I error rate. Given the negative consequences of a false negative (i.e., categorizing as “no change” when in fact there is change) in clinical practice, it is unreasonable to hold a Type I error rate) that is too rigorous.

### C. Standardized Pre-Post Mean Change (Cohen’s *d*)

Cohen’s *d* is a preferred effect size statistic for illustrating change, especially from pretest (Becker, 1988; Morris, 2000). Specifically:

$$d = \left(1 - \frac{3}{4n - 5}\right) \frac{mean_y - mean_{pre}}{SD_{pre}} \quad (5)$$

Here,  $mean_{pre}$  ( $SD_{pre}$ ) are the mean and standard deviation at pretest,  $mean_y$  is the mean at a subsequent point in time (e.g., posttest), and  $n$  is the total sample size. The reason why the standard deviation of the pretest is used rather than the pooled standard deviation is because the pretest standard deviation is not affected by the treatment or other effects that could be confounding at subsequent measurements (Becker, 1988).

A positive change in the magnitude of  $d = 1.0$  indicates that the patients, on average, have improved to where the average is now at the 85th percentile as compared to when they started treatment. Cohen (1988) tentatively defined effect sizes as  $d = 0.2$  (small),  $d = 0.5$  (large), and  $d = 0.8$  (large), with a caveat that "there is an inherent certain risk in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (p. 25). Indeed, despite the hesitancy with which he proposed these qualitative descriptors for effect size ranges, these remain probably the most often cited criteria for evaluating effect size. However, as recent work on benchmarking reveals, different types of questionnaires are likely to result in different effect sizes. For example, clinician-completed measures tend to have larger effect sizes than patients’ self report measures; likewise, disorder-specific measures, which focus on a discrete set of symptoms, tend to show larger effect sizes than more global measures of distress (Minami, Wampold, Serlin, Kircher, & Brown, 2007). The ACORN BAOQ would obviously be classified as a client self-report global symptom measure.

As it happens, it appears that Cohen’s suggestion for describing effect size magnitude fits well with results from psychotherapy research. Minami et al.’s (2007) meta-analysis of outcomes for clinical trials of psychotherapy for depression utilizing patients’ self report global distress measures arrived at an average effect size of 0.79 (large by Cohen’s criteria).<sup>1</sup> Therefore, independent of Cohen’s criterion, the magnitude of  $d = 0.8$  is a good criterion to use for psychotherapy.

#### D. Residuals

Any observed change is only an estimate of the “true” change. In addition to random error (as described above as  $SE$ ), other factors that are known to influence treatment outcomes must be taken into consideration and modeled.

Among the different predictors, initial severity is known to affect the magnitude of change in subsequent assessments. In part due to regression to the mean, higher the initial score the higher the likelihood that subsequent assessments would be closer to the mean. Another known predictor in psychotherapy is the time between the two assessments. When clients are in treatment at regular intervals, this is roughly interchangeable with number of sessions. Clients’ age, sex, and clinical diagnosis also can be predictors, although there has been no reliable evidence that treatment outcomes are significantly affected by age, sex, and clinical diagnosis.

When properly modeled, the general linear model allows for calculations of predicted scores and residuals. In the context of treatment outcome, the predicted scores are the expected degree of change based on what is modeled (e.g., initial severity, length of treatment). The residuals, on the other hand, is the difference between the predicted score and the observed degree of change. When larger values are in the direction of better outcomes, greater residuals indicate more positive outcome than predicted by the model. The residualized score at the last assessment (e.g., posttreatment) is referred to as the “Change Index Score.”

#### ACORN Criteria for Effectiveness (ACE): Example

Table 1 presents the base statistics for the ACORN BAOQ. The  $SE$ , RCI, and change score are converted to Cohen’s  $d$  for comparison in standardized units.<sup>2</sup>

Table 1: ACORN BAOQ Base Statistics

	Raw Scores	$d$
Intake $M$ ( $SD$ )	2.40 (0.52)	
Post $M$	1.95	
Alpha Coefficient	.91	
$SE$	0.16	0.30
RCI	0.43	0.83
Change Score	0.45	0.87

Based on the above information, there are two methods that are currently conceptualized that would qualitatively illustrate the clinician’s caseload, notably, change criteria categorization and

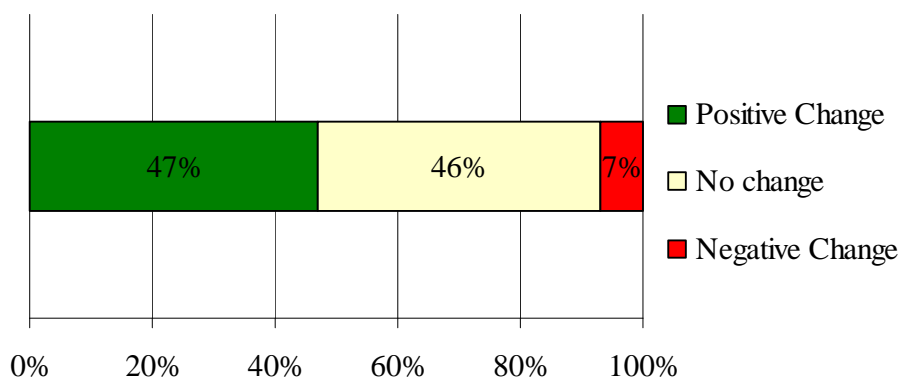
benchmarking. The first method classifies patients within each clinician based on change criteria using the RCI and *SE*, whereas the second method provides an overall assessment of the clinician’s effectiveness. Both methods fully take into consideration differences in case mix among clinicians.

#### A. Change Criteria Categorization

In the ACORN database, approximately 47% of the cases were classified as “positive change” and 7% of the cases were classified as “negative change” based on the RCI. This leads to a three-level classification of change as in Chart 1.

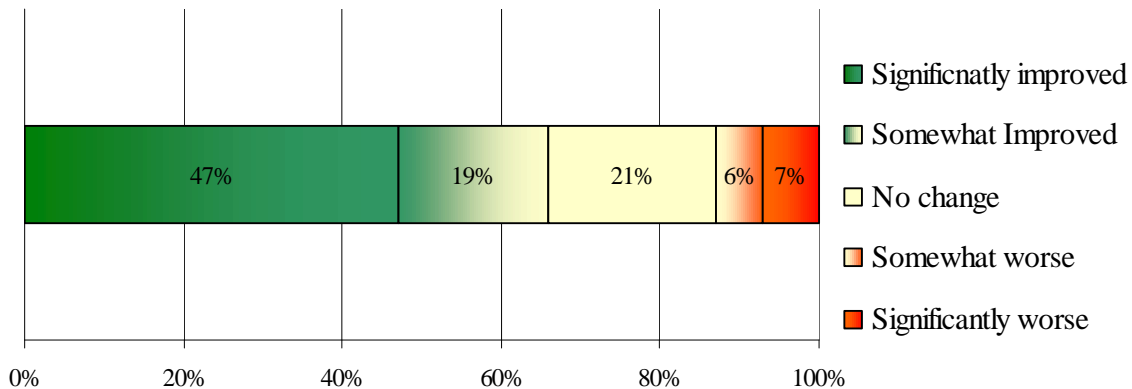
As previously noted, the RCI is a conservative criterion, leaving approximately 46% of the patients as “unchanged.” Given that the average pre-post effect size is approximately  $d = 0.83$ , which is comparable to effects observed in clinical trials, moderate changes that are likely meaningful for the patients are not captured using the RCI. For example, while Cohen has classified an effect size of  $d = 0.5$  as a medium effect, the RCI would classify this as “unchanged.”

Chart 1: Three-Level Change Classification Using the RCI



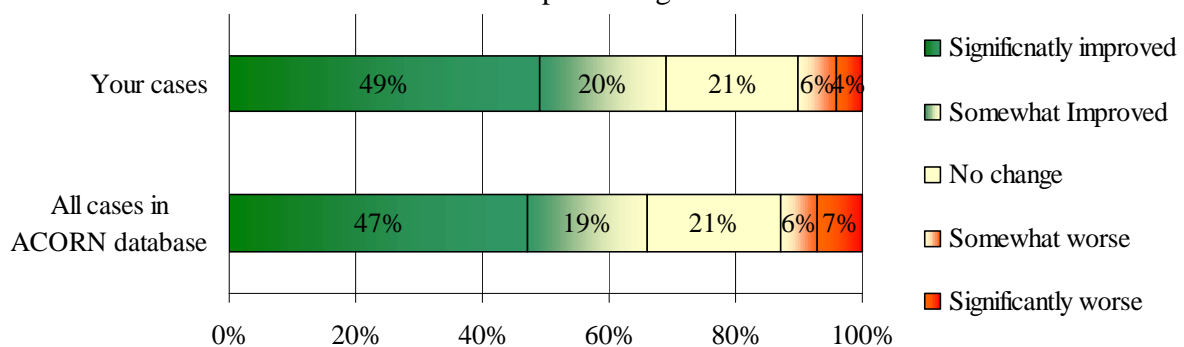
While it is difficult to quantitatively determine the minimum amount of change that could be considered clinically meaningful, Eisen et al. (2007) support the use of *SE* as threshold for clinically significant change. The *SE* for the ACORN BAOQ is  $d = 0.30$ , which is larger than what Cohen classifies as small (i.e.,  $d = 0.2$ ). Therefore, in addition to the RCI, it is likely meaningful to use the *SE* as additional criterion to classify meaningful change. This leads to the five-level change classification as in Chart 2.

Chart 2: Five-Level Change Classification Using the *SE* and RCI



As could be seen in comparison to Chart 1, “no change” is further divided into 3 parts based on the *SE*. The cases between the RCI and *SE* are now classified as “somewhat improved” or “somewhat worse,” based on the direction of the residual. The five-tiered method for describing change permits a more nuanced presentation of the results, which likely better reflect clinician’s perceptions of improvement. In addition, as additional 19% exceed the *SE*, this method allows for classification of 66% of the caseload to be classified as “improved” as compared to 47%. Chart 3 illustrates how the above could be used in comparison to the total ACORN database.

Chart 3: Individual Clinician Outcomes Comparison against ACORN Database



To clinicians, the charts intuitively provide information as to where their outcomes stand in relation to all cases in the database. Note that the classifications are all residualized, i.e., differences in initial severity, population, etc., are already taken into consideration), and thus, the ACORN database categorizations provide clinicians with a relative (i.e., compared internally within the ACORN database rather than “absolute”) and unbiased benchmark of their caseload.

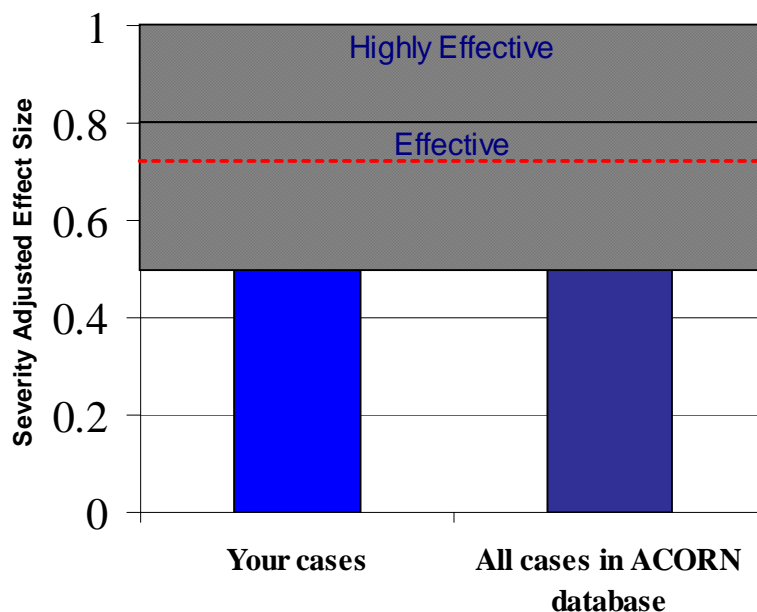
### B. Caseload Benchmarking

However, it is also important for the individual clinician to know their effectiveness in comparison to a more “absolute” benchmark. For example, what does being approximately at par with the average observed in the ACORN database mean? While there is no definite way to quantify effectiveness, one method is to compare against the most rigorous criteria that is currently available—against published clinical trials.

As previously mentioned, Minami et al. (2007) derived benchmarks from published clinical trials of adult depression.<sup>3</sup> With client self-report of global clinical symptoms (as the ACORN BAOQ), the magnitude of pre-post effect was estimated at  $d = 0.79$ . As this is approximately Cohen’s guideline for a large effect size (i.e.,  $d = 0.8$ ), it seems reasonable to utilize his suggestion in considering what constitutes a “highly effective” average outcome, adjusted for case mix. Also using Cohen’s other guidelines, an average effect size of over  $d = 0.5$  could be constituted as “effective.” In addition, Minami et al. (2008) defined a lower band of within 10% below the clinical trials benchmark as a clinical equivalence (i.e., effect sizes over  $d = 0.72$ ). In other words, if a clinician’s average effect size is within this band, their outcomes could be considered practically equivalent to the treatment effect observed in clinical trials.

Chart 4 is an example using these criteria to provide an “absolute” (rather than relative to others in the ACORN database) qualitative interpretation of a clinician’s average observed effect size as well as that of the ACORN database against clinical trials. Note that for both the individual clinician and the ACORN database, the effect sizes are adjusted for case mix and thus are not “raw” observed values. Effect sizes above  $d = 0.8$  is considered highly effective and those above  $d = 0.5$  is considered effective. The pink dotted line illustrates the minimum criterion ( $d = 0.72$ ) to interpret one’s clinical outcomes as at par with clinical trials.

Chart 4: Individual Clinician Caseload Benchmarking against Clinical Trials



### C. Adjusting for Sample Sizes and Measurement Error

Any effort to benchmark therapists must account for the effects of small sample sizes on measurement error. Measurement error can be minimized by assisting on large sample sizes, but this may be the unfortunate effect of making it more difficult to identify effective clinicians.

Hierarchical Linear Modeling (HLM) can be used to estimate a confidence interval around the estimate of SAES. An 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 & Serlin 2000).

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 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. 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. In these instances, use of a 90% confidence level may be justified.

The question of setting a minimum sample size necessary to calculate SAES is to some degree arbitrary because this is a new development in research that have not reached consensus among researchers. From a practical standpoint, the use of random-effects HLM permits the use of relatively small sample sizes because of the above conservativeness. For the reason, a minimum sample size of at least 10 cases is recommended, while 15 or more cases is preferable.

#### *D. Consistency of data collection*

One obvious threat to the integrity of the outcome data is any tendency of the clinician to selectively choose which patients to which administer questionnaires. A low percentage of cases with multiple assessments can indicate that a clinician is being selective (though most likely a simple matter of inconsistency) in the use of the questionnaires.

Analyses of large samples from outpatient treatment settings indicate that well over than 50% of cases have enough sessions so that multiple assessments should be expected. For this reason, it is recommended that at least 50% of cases with intake scores in the clinical reason have repeated measures in treatment before the SAES is taken as strong practice based evidence of the clinician's effectiveness.

## Summary

This document is a methodological proposal for qualitatively evaluating a clinician's caseload, utilizing what is currently available from classical psychometric theory and benchmarking. Obtained assessments are adjusted for case mix so as comparisons are as unbiased as possible, and the adjusted effect sizes are quantitatively compared against "relative" (i.e., with ACORN database) and "absolute" (against clinical trials) benchmarks.

Use of random-effects HLM assures that clinicians are given the "benefit of the doubt" when sample sizes are low. Based on our analyses of large samples of clinicians, we estimate that use of a minimum sample size of 15 and a 90% confidence level that the SAES is  $d = 0.5$  or higher, at least 80% of clinicians will be classified as effective provided that they have collected multiple assessments on at least 50% of cases in the clinical range.

### *Footnote*

<sup>1</sup>Please note, as in Minami et al. (2007), that pre-post standardized mean change is significantly affected by the sensitivity and reactivity of the outcome measures. Therefore, the use of Cohen's guideline for interpreting the magnitude of effect should not be applied with measures other than client self-report of global symptom measures.

<sup>2</sup>Technically, the values in Table 1 are in  $g$ , not  $d$ , as it does not include the correction factor (i.e.,  $1 - 3/4(n - 1)$ ) that becomes negligent in large sample sizes. However, the document uses  $d$  for consistency.

<sup>3</sup>Note that the benchmarks were derived for adults diagnosed with major depressive disorder using the DSM. The residuals are thus adjusted for diagnosis.

## References

- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112 (1), 155-159
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum
- Eisen, S.V., Ranganathan, G., Seal, P. & Spiro, A. (2007). Measuring Clinically Meaningful Change Following Mental Health Treatment. *Journal of Health Services and Research*. 34, 272-288.
- Martindale, C. (1978) The therapist-as-fixed-effect fallacy in psychotherapy research. *J. Consult. Clin. Psychol.* **46**, 1526-1530.
- Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (In press). Using clinical trials to benchmark effects produced in clinical practice, *Quality and Quantity*
- Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007a). Benchmarks for psychotherapy efficacy in adult major depression, *Journal of Consulting and Clinical Psychology*, 75, 232-243.
- Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E., Brown, G. S., & Kircher, J. (2007b). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, 40, 73-83.
- Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199-228). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Serlin, R.C., Wampold, B.E., Levin, J.R. (2003) 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.
- Wampold, B.E., Serlin, R.C. (2000) The consequences of ignoring a nested factor on measures of effect size in analysis of variance designs. *Psychol. Methods*, **4**, 425-433.