Psychotherapy Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tpsr20

Effect size calculations for the clinician: Methods and comparability

Jason A. Seidel[a], Scott D. Miller[b] & Daryl L. Chow[b]
[a] Colorado Center for Clinical Excellence, Denver, CO, USA
[b] International Center for Clinical Excellence, Chicago, IL, USA
Published online: 05 Nov 2013.

To cite this article: Jason A. Seidel, Scott D. Miller & Daryl L. Chow , Psychotherapy Research (2013): Effect size calculations for the clinician: Methods and comparability, Psychotherapy Research, DOI: 10.1080/10503307.2013.840812

To link to this article: http://dx.doi.org/10.1080/10503307.2013.840812

Psychotherapy Research, 2013
http://dx.doi.org/10.1080/10503307.2013.840812

Routledge
Taylor & Francis Group

EMPIRICAL PAPER

Effect size calculations for the clinician: Methods and comparability

JASON A. SEIDEL[1], SCOTT D. MILLER[2], & DARYL L. CHOW[2]

[1]*Colorado Center for Clinical Excellence, Denver, CO, USA* & [2]*International Center for Clinical Excellence, Chicago, IL, USA*

(Received 9 May 2013; revised 5 July 2013; accepted 30 August 2013)

Abstract
Objective: The measurement of clinical change via single-group pre-post effect size has become increasingly common in psychotherapy settings that collect practice-based evidence and engage in feedback-informed treatment. Different methods of calculating effect size for the same sample of clients and the same measure can lead to wide-ranging results, reducing interpretability. **Method:** Effect sizes from therapists—including those drawn from a large web-based database of practicing clinicians—were calculated using nine different methods. **Results:** The resulting effect sizes varied significantly depending on the method employed. Differences between measurement methods routinely exceeded 0.40 for individual therapists. **Conclusions:** Three methods for calculating effect sizes are recommended for moderating these differences, including two equations that show promise as valid and practical methods for use by clinicians in professional practice.

Keywords: psychotherapy outcomes; effect size; effectiveness; clinical change; outcome measurement

Methods for measuring clinical change in real-world psychotherapy settings have proliferated in recent years along with software and web-based platforms for sophisticated outcomes analysis (e.g., Client Voice Innovations, 2013; CORE System Trust, 2013; Ed & Psych Associates, 2011; FIT-Outcomes ApS, 2013; Health Factors, 2013; Miller & Duncan, 2004; OQ Measures, 2013; Psytek Ltd., 2005; Seidel & Miller, 2012; Wiarda Group, 2013). Increasingly, clinicians and agencies use outcome instruments that measure general subjective well-being (SWB) as a broad index of clinical change. While its construct validity is open to debate (e.g., Eid & Larsen, 2008), SWB provides a feasible way for psychotherapists to track general clinical change especially since diagnostic comorbidity occurs among more than 50% of clients in a typical clinical caseload (Rodriguez et al., 2004). Moreover, few clients and therapists are willing to complete and score (on a session-by-session basis) longer symptom-based measures that cover a more comprehensive array of diagnoses or problem sets when they are as long as 40 items and take longer than 5 minutes (Australian Mental Health Outcomes and Classification Network, 2005; Brown, Dreis, & Nace, 1999; Duncan, 2012; Miller, Duncan, Brown, Sparks, & Claud, 2003).

Many clinicians and mental health agencies analyzing such practice-based evidence (e.g., Andrews, Twigg, Minami, & Johnson, 2011; Barkham et al., 2001; Barkham, Mellor-Clark, Connell, & Cahill, 2006; Borkovec, Echemendia, Ragusea, & Ruiz, 2001; McMillen, Lenze, Hawley, & Osborne, 2009) do so through a single-group pretest-posttest (SGPP) design, computing a repeated-measures (pre-post) effect size (*ES*) as a metric of change. Notwithstanding the view that *ES* can serve as a "scale-free" score allowing comparisons between different instruments (e.g., Vacha-Haase & Thompson, 2004), research shows that different sampling methods, instrument sensitivities, and calculation methods can yield different outcomes when measuring the same presumed construct (e.g., Becker, 1988; Dunlap, Cortina, Vaslow, & Burke, 1996; Morris, 2008; Morris & DeShon, 2002; Olejnik & Algina, 2003; Ray & Shadish, 1996; Wampold & Brown, 2005; Werbart, Levin, Andersson, & Sandell, 2013).

Correspondence concerning this article should be addressed to Jason A. Seidel. Colorado Center for Clinical Excellence, 1720 S. Bellaire St., Suite 204, Denver, 80222 USA. Email: jseidel@thecoloradocenter.com

As evidence mounts that psychotherapy effectiveness can be augmented through outcome and alliance monitoring and client engagement in that monitoring (Baldwin, Wampold, & Imel, 2007; Lambert et al., 2002; Whipple et al., 2003), and that effectiveness also can vary a great deal among psychotherapists (Anderson, Ogles, Patterson, Lambert, & Vermeersch, 2009; Crits-Christoph & Mintz, 1991; Kim, Wampold, & Bolt, 2006; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi et al., 2006; Serlin, Wampold, & Levin, 2003; Wampold & Brown, 2005; however, cf. Crits-Christoph & Gallop, 2006), the need for valid and practical methods for analyzing and comparing individual client change and outcome differences among therapists becomes clearer (Miller, Hubble, Chow, & Seidel, 2013). Measuring clinical change "on the fly" allows clinicians and agencies to improve service delivery with the client in the room, and more appropriately direct ongoing professional development resources. While the simplest *ES* formulas can be calculated with basic math skills and a hand calculator, and more complex *ES* calculations can be performed on spreadsheet programs such as Microsoft Excel, a lack of clarity remains about the effect of different calculation methods on the obtained *ES*.

ES formulas that some clinician-researchers have used (e.g., Cohen's *d*, Glass's Δ, and Hedges' *g*) measure differences between two independent groups, not the difference most clinicians need and want to measure (i.e., within a single group). Many of the more complex *ES* algorithms for interpreting clinical change in SGPP samples either are held as proprietary trade secrets by outcomes software publishers or are too complex for most clinician-researchers to understand or implement. In other words, these complex methods lack either transparency, feasibility, or both. Inadequate disclosure of proprietary algorithms can prevent methodological scrutiny of assumptions and potential error, decreasing the explanatory power that statistical procedures are meant to provide. While confidence intervals can be calculated for *ES* to increase interpretability, their usefulness is reduced by the likelihood of wide intervals that defy practical interpretation (at commonly used significance levels, and for the sample sizes typically collected by clinician-researchers; Werbart et al., 2013), and also by their conceptual difficulties (e.g., frequentist versus Bayesian methods and interpretations; Crawford, Garthwaite, & Porter, 2010). Outside academia, such matters reduce feasibility. When almost no mental health service settings employ the services of highly specialized outcomes statisticians (and outcomes software publishers have not resolved these difficulties), a gap is created between the ideal rigor with which researchers might want outcomes to be analyzed, and the practical question of whether outcomes analysis takes place at all.

The principle of statistical parsimony argues for a balance between simplicity and accuracy, requiring that explanations and models be "good enough" but also accessible to a reasonable number of people (May, 2004). For example, Becker (1988) recommended a complex equation to correct for the bias in a simple equation for pre-post *ES*. Yet, Hedges and Olkin (1985) showed that this additional step is of little practical consequence for sample sizes greater than 20. For clinicians, mental health administrators, and consumers to understand and properly account for clinical effectiveness data, the statistical models should be comprehensible to as many people as possible, making them more open to critique and refinement. Moreover, clinicians should understand how these different methods change the *ES*, and how to select a method that most accurately characterizes clinical change with the least amount of systematic bias and error.

Types of Single-Group Pre-Post Effect Sizes

*ES*s typically used in SGPP designs by clinician-researchers are fractions, with a change score (e.g., posttest minus pretest) as the numerator and sample variability (e.g., standard deviation) as the denominator:

$$ES = \left(M_{\text{post}} - M_{\text{pre}}\right)/SD \tag{1}$$

Adjustments to either the numerator or denominator (or both) address various validity concerns in *ES* calculation. *ES*s that are calculated from a sample's raw pretest and posttest data—without any adjustments or corrections to the clinical change represented in the numerator—can be called a "raw" *ES*. Raw *ES*s from an identical dataset may have different values depending on the denominator used to standardize the clinical change score. *ES*s calculated from a sample's adjusted pretest and posttest data—such as change scores that are first transformed by case-mix prediction models and then divided by sample variability—can be called "adjusted" *ES*s.

Alternate versions of raw *ES* are obtained from the use of different variability terms in the denominator. An *ES* based on pretreatment variability (ES_{pre}) is the change from pretreatment to posttreatment divided by the standard deviation of the sample's pretreatment scores (SD_{pre}; Becker, 1988). The SD_{pre} standardizes the change in well-being based on the dispersion of initial distress scores among the people in the sample-of-interest. A high pre-post difference among clients in the sample who have

little variability (i.e., low SD_{pre}) among their pretreatment scores will yield a high ES_{pre}; while an identically high pre-post difference in clients who have more variability in pretreatment distress (i.e., high SD_{pre}) will yield a lower ES_{pre}. Therefore, a drawback of using the SD_{pre} as the divisor is that therapists with clients expressing a wider range of distress at intake are "penalized" by lower ES_{pre} scores; and therapists whose clients show a tighter range of distress at intake are "rewarded" with higher ES_{pre} scores (see Table I). On the other hand, and as shown below, all *ES* calculation methods—from the simplest to the most complex—have shortcomings; and the ease of computation and interpretation makes ES_{pre} a good first step for understanding and measuring clinical change.

Using the *SD* of pre-post differences (SD_{diff}, i.e., the *SD* of the change in distress, rather than the *SD* of the pretreatment distress) shifts the focus of standardization from the sample's intake variability to the variability in how much change clients experience over the course of their therapy. As with the ES_{pre}, the ES_{diff} allows systematic differences between samples or clinicians to affect the change score in ways that may not be appropriate (Table II). For example, given identical pretreatment-score variability and identical averages for pretreatment well-being and posttreatment well-being, a clinician whose clients end treatment with more variable posttreatment scores will have a lower ES_{diff} than a clinician with a tighter clustering of either high or low scores at posttreatment. Therapists whose clients report substantial changes but who have a broader dispersion of change scores are therefore penalized in comparison with therapists whose clients report substantial changes with a tighter dispersion of change scores, and also in comparison with therapists whose clients report less change but with a tighter dispersion of change scores.

Both the ES_{pre} and ES_{diff} are impacted by between-therapist variables (pretreatment-distress variability and change variability, respectively) that may bias what a standardized-change *ES* is intended to estimate: the difference between two averages as a function of the variability in an instrument's scores within a particular population-of-interest (e.g., intensive-outpatient psychotherapy clients, divorced parents, Latino teenagers, or 50–65-year-old US adults).

Table I. Effect of pretreatment score variability on ES_{pre}

Pretreatment score variability (*SD*)	M_{pre}	M_{post}	SD_{pre}	ES_{pre}
High	17	27	10	1.00
Low	17	27	6	1.67

Note. ES_{pre} = effect size based on the standard deviation of pretreatment scores; M_{pre} = mean of pretreatment scores; M_{post} = mean of posttreatment scores; SD_{pre} = standard deviation of pretreatment scores.

Table II. Effect of posttreatment score variability on ES_{diff}

Pretreatment Score	Posttreatment Score	Difference	SD_{diff}	ES_{diff}
Therapist with higher variability of posttreatment scores				
10	29	19		
12	24	12		
14	32	18	10.5	.95
22	15	−7		
27	35	8		
Therapist with posttreatment high-score clustering				
10	28	18		
12	22	10		
14	30	16	7.2	1.39
22	27	5		
27	28	1		
Therapist with posttreatment low-score clustering				
10	25	15		
12	26	14		
14	32	18	8.1	1.24
22	26	4		
27	26	−1		

Note. ES_{diff} = effect size based on the standard deviation of difference scores.

A third method of *ES* calculation (ES_{ref}) involves selecting a reference-based "standard deviation of the instrument," an SD_{ref} that "fixes" an approximate variability of a population-of-interest and characterizes the pretreatment score variability based on that particular reference group, treatment setting, etc. If there are reference groups available that show a relatively small range of *SD* values for a given population (though a method for defining and operationalizing an "acceptable" range or confidence interval of *SD*s on SWB instruments remains to be determined), then this SD_{ref} might be said to be a better divisor for the standardization of an obtained change score than the SD_{pre} or SD_{diff} because it does not penalize or reward an individual clinician's *ES* based on comparative variability of their own sample's pretreatment distress or change scores (e.g., Miller, Duncan, Sorrell, Brown, & Chalk, 2006, used the SD_{pre} of a previously published nonclinical sample as an SD_{ref} for their treatment sample "as an indication of how much clients in the [current] study improved relative to [the variation of] a normal population"; p. 11).

While an ES_{ref} reduces variation in the error term used for comparison between clinicians or between a clinician and benchmark, it does not address the questions that may arise about why a particular clinician's or agency's SD_{pre} is substantially different from an SD_{ref}, or whether a particular reference group (or multiple groups) from which an SD_{ref} was

chosen is demographically similar to (or used measurement methods that might affect the *SD* for) a given sample. As one example, SD_{ref} values for the outcome rating scale (ORS; Miller et al., 2003) can vary markedly between large samples. So while the ES_{ref} might be considered an improvement on the ES_{pre} or ES_{diff}, external validity issues and decision rules for its use currently complicate the interpretation of ES_{ref}.

An alternative to the raw *ES* formulations (which use different divisors based on pretreatment, difference-score, or reference-group *SD*) is a *t* value converted to Cohen's *d*, i.e., a conversion of the magnitude of difference between groups yielded by a repeated-measures *t* test to an *ES*. Dunlap et al. (1996) introduced a corrected formula for repeated measures (ES_{RMC}) because the standard formula for translating a *t* value into an *ES* was based on an independent-groups *t* value (ES_{ri}) that was uncorrected for the reduced random error in a repeated measures *t* test. Using the standard (independent-groups) formula for converting *t* to *d* for a repeated-measures sample (which, in that instance, can be called ES_{RMU}: an uncorrected ES_{RM}) without adjusting for the difference in degrees of freedom between the two types of *t* test will overestimate the resulting *ES*. Dunlap et al. explained that this overestimation of *ES* is due to the significant correlation between repeated-measures scores in an SGPP design that increases power through a reduction in the standard error of the difference in scores. This overestimation is a concern with any *ES* calculations that are derived from a correlation coefficient (e.g., from either *t* or *r* statistics). Dunlap et al.'s correction also prevents larger sample sizes (which would increase the *t* value) from increasing the resulting *d*; but Dunlap et al. did not report a comparison between this correlated-score method of *ES* calculation (which uses SD_{diff} in the denominator) and raw-*ES* methods that do not account for pre-post correlation. As shown below, the ES_{RMC} will yield a different *ES* than those generated by raw *ES* calculations.

Finally, a *case-mix adjusted* or *severity-adjusted ES* (ES_{SA}) takes into account the considerable effect of factors such as pretreatment distress as predictors of clinical change. Pretreatment distress is a highly robust and consistent predictor of outcomes in psychotherapy research and is almost self-evident: the better one feels when starting therapy, the better one is likely to feel by the end of therapy; but the worse one feels at the start, the greater the expected degree of improvement (Wampold & Brown, 2005). Adjusting the *ES* for "expected change" based on pretreatment distress allows the variability in client severity that may exist at intake among different agencies or therapists to be taken into account when measuring *ES*.

With the ES_{SA}, the same absolute change in score is weighted more positively for clients who already start with high well-being (when improvement from high well-being to even higher well-being has been relatively small in a large reference sample), and weighted less positively for clients who start with low well-being (when improvement from low well-being has tended to be large). Some authors (e.g., Crits-Christoph & Gallop, 2006; Okiishi et al., 2006) have argued that severity-adjusted analyses of change also ought to include time or number of sessions (e.g., through hierarchical linear modeling) because the number of sessions provides additional information about treatment efficiency or the relative speed of change between therapists. For example, Crits-Christoph and Gallop argued that endpoint (and by extension, pre-post change) analysis "disregards the time effect and treats each endpoint as if it was obtained at the same point in time and assumes that patients would have no change beyond that endpoint" (2006, p. 178). However, neither of these assumptions follows from the use of endpoint or pre-post change analysis. The latter take advantage of the reality of how clients use and end their treatment: when they feel that there is not much further to be gained by continuing (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009; Stiles, 2013).

As a measure of the difference between the client's well-being when the client felt distressed enough to start therapy and when the client (presumably) felt that little would come of continuing, ES_{SA} based on a simple change-score analysis (rather than HLM) makes good sense. This kind of analysis does not assume "patients would have no change beyond that endpoint"; rather that the endpoint can be assumed to roughly match when clients perceive that further sessions will add little additional benefit to whatever change they have experienced or are likely to continue experiencing with or without this particular therapist. While it may be important to know whether certain therapists are more efficient in how quickly their clients experience "good enough" clinical change from session to session (and ought to receive more statistical credit based on faster clinical change in their samples), in real-world practice clients typically self-regulate: as they feel better, they end treatment (Baldwin et al., 2009; Leichsenring & Rabung, 2008) and as a result, there is often little correlation between time and effectiveness when treatment completers are analyzed (Werbart et al., 2013). Often, clients return several months or years later, whether for a brief "check in" session or multiple episodes of therapy. Measuring therapists' average rates of change per session or per

week can create a misleadingly narrow expectation of the trajectory of individual change. On an individual basis, changes in well-being often vary dramatically from session to session; and episodes of treatment also show markedly idiosyncratic patterns-of-attendance and rates of change (Baldwin et al., 2009; Hafkenscheid, Duncan, & Miller, 2010).

Baldwin et al. (2009) showed how an emphasis on the overall rate of change can conceal the unpredictable and self-regulatory nature of session-by-session psychotherapy while a change score simply disregards the in-process trajectories and focuses attention on degree of change by the end of treatment. Further, Wang and Duan (in preparation) showed that in contrast to the statistical noise that the time factor adds to a time-series or longitudinal analysis, the simplicity and statistical power provided by endpoint and change-score analysis can be more appropriate.

However, while the ES_{SA} can improve the relative accuracy of measuring standardized change by adjusting the effect size based on each client's level of pretreatment distress, it does not address the problem of appropriate instrument selection. The ES_{SA} still may not capture the true range that would be obtained from an instrument with a different floor, ceiling, or sensitivity. Also, a counter-argument to using a severity adjustment based on pretreatment distress is that if therapists with very different clientele distress are compared, the ES_{SA} will (relatively) "reward" therapists if they help their mildly distressed clients improve dramatically, but "punish" therapists if they help their highly distressed clients improve dramatically, simply because the likelihood of these clients' respective improvements is different. From a public health perspective, the latter client group might be of greater concern even if more likely to show the improvement. Moreover, calculating the ES_{SA} still entails the problem of which *SD* to use (e.g., SD_{pre}, SD_{diff}, or SD_{ref}) as the basis for standardizing each client's change. For example, Miller et al. (2006) used the SD_{ref} of a small (n = 86) nonclinical reference group made up of counseling center graduate students, faculty, and staff, rather than the SD_{pre} of the large sample (n = 6424) for which the ES_{SA} was calculated. Since publication of Miller et al. (2006), SD_{pre} in published accounts of clinical samples using the ORS have ranged widely (between 6.5 and 10.4; Reese, Norsworthy, & Rowlands, 2009).

Finally, Werbart et al. (2013) showed the wide range of obtained *ES* using the same calculation methodology but with three different well-being instruments for clients receiving three different types of therapy. Even when narrowing the analysis to measure change only for treatment completers within each treatment type, marked differences were found in *ES* (routinely exceeding .40) between the different measures of well-being. And *ES* confidence intervals exceeded .90 in over half of the nine subsamples, making the intervals of no practical use to clinicians, supervisors, or clients. Considering the range of methods used for measuring *ES*, the aim of the current study is to systematically compare these methods in a sample of therapists to clarify the relationships between them and offer a clearer way forward for practicing clinicians who want to provide feasible, transparent, and accurate reporting of their effectiveness. While this does not address the separate concern of the way different SWB scales may affect outcomes, a more unified analytic method might simplify those comparisons as well.

Method

Participants

Participants came from two pools: a web-based outcome management system, and the first author's database of independent psychologists who had submitted their outcome data for detailed analysis. The web-based database contained a large sample of psychotherapists in a variety of undefined and anonymous international treatment settings all of whom administered a web-based version of the ORS to their clients. The independent psychologists were a small sample-of-convenience using the ORS in their private practices on three separate continents.

Inclusion criteria. For the large sample, 38,608 clients with at least a pretreatment score and a second score were reduced to a subsample of those clients of therapists who had between 30 and 500 cases in the database. The last available score for each client was used as the second score for the purpose of pre-post analysis. The resulting 17,285 clients seen by 262 therapists were then subject to the exclusion criteria below. Three therapists (Therapists A, B, and C) from this large sample were selected randomly as exemplars from those with n > 99 clients. An additional group of three psychologists in private practice with n > 99 (Therapists D, E, and F) were self-selected members (from the USA, Australia, and The Netherlands, respectively) of an international organization of outcomes-oriented psychotherapists and mental health administrators who had submitted their ORS data for detailed analysis. These three private practitioners used either the paper-and-pencil version (Therapists D and E) or a combination of two different computer-based and web-based software products (Therapist F). A follow-up sample of seven additional therapists (Therapists G–M, each

with $n > 50$) was selected to determine whether the relationships among *ES* values for Therapists A–F would be replicated. From the large web-based database, six therapists with stratified outcomes were chosen (two with low ES_{pre} [< .30] called Therapists G and H, two with midrange ES_{pre} [.60–.70] called Therapists I and J, and two with high ES_{pre} [> 1.00] called Therapists K and L). In addition, data from an additional US psychologist in private practice (Therapist M) were analyzed in this follow-up sample.

Exclusion criteria. Therapists with caseloads greater than 500 were excluded from the large web-based database as highly atypical (given the recent advent of the web-based program). Therapists with caseloads smaller than 30 were excluded as more likely to: (a) need a correction for small *n*; (b) have a less reliable *ES*; (c) have a high proportion of open cases, increasing the proportion of last available scores that were not posttreatment scores; and (d) be influenced by artificial or practice data as clinicians learned a new software system without the oversight of researchers vigilant against inaccurate data entry. Also, therapists in this subsample who had an unusually high proportion of clients with "perfect" well-being at intake—defined as ≥ 5% of their clients having ORS total scores of 40—were excluded. A large proportion of perfect well-being scores at the first session indicates a different kind of clientele than is seen in typical practice, e.g., clients in mandated treatment. The incidence of clients reporting perfect well-being in their first session was low (1.8% of the 17,285 clients in the sample). However, the incidence was not uniformly distributed: 40.5% of the 262 therapists had one or more such clients, and of those 106 therapists who did have at least one of these clients, 68.9% of them had < 5% of their clients reporting perfect well-being at intake. The mean (*SD*) percentage of clients reporting perfect well-being at intake among the 262 therapist caseloads was 1.8% (3.6%). The mean (*SD*) percentage of clients reporting perfect well-being at intake was 4.5% (4.4%) among the 106 therapist caseloads with at least one such client. Therefore, the exclusion of therapists with ≥ 5% of their caseload reporting perfect well-being at intake, while somewhat arbitrary, falls between the mean + 1.0SD of the total incidence rate among the 262 therapists (i.e., 5.4%) and the mean of the incidence rate among the 106 therapists who had at least one such client in their caseload (i.e., 4.5%). The exclusion of the 33 therapists with ≥ 5% of these perfect-well-being clients yielded a final subsample of 15,398 clients seen by 229 clinicians, or 87.4% of the 262 therapists. No demographic or other contextual information (e.g., age, ethnicity, gender, diagnostic category, clinical setting, or professional experience) except treatment duration was available for therapists or clients.

Measure

Outcome Rating Scale (ORS). The ORS is a four-item self-report SWB instrument that is available either as a free (for individual use) paper-and-pencil form, or through several software programs. The scale was derived from the Outcome Questionnaire-45.2 (OQ-45.2; Lambert et al., 2004) as an ultra-brief alternative that could more feasibly be given at every session without overburdening the client or clinician with the administration and scoring process, and that would appeal to a broader range of clientele, including those with less education (Miller et al., 2013). Three well-being items (individual, close relationships, and work/school/friendship) mirror the three subscales of the OQ-45.2, and there also is a fourth "overall" well-being item. The format is a numberless visual-analogue scale with a 10 cm horizontal line segment under each item that clients are instructed to mark with a vertical stroke to correspond with "how well you have been feeling" in the past week, with marks to the left representing "low levels" and marks to the right representing "high levels."

While not apparent from the lack of numbers, the scale is set up as an interval scale with the left-most point set at 0 cm. The clinician then measures with a ruler to obtain the score to the nearest millimeter for each item, and sums the four scores for a total SWB score. The ORS takes about 30–60 seconds to administer, depending on the engagement of the client, and another 30–60 seconds to score. The ORS is designed to be administered at each session, with total scores plotted on a line graph and presented to the client for possible discussion about changes in SWB over the course of treatment. The software versions use a graphical-user-interface "slider" that the client manipulates with a mouse or finger-on-touchscreen, and item and total scores are calculated automatically.

Compliance rates for session-by-session administration in routine practice are reported as being high in comparison with longer instruments (Brown, 2006; Miller et al., 2003). Correlation with OQ-45.2 subscales and total score (e.g., −.69, −.74, −.59 for total score) is adequate (Bringhurst, Watson, Miller, & Duncan, 2006; Campbell & Hemsley, 2009; Miller et al., 2003), and internal consistency is high (e.g., .91, .90, .87; Bringhurst et al., 2006; Campbell & Hemsley, 2009; Miller et al., 2003). Test-retest reliability is moderate in nonclinical

samples (e.g., .80 and .58 for an approximately 1–2-week delay; Bringhurst et al., 2006; Miller et al., 2003), and this is expectable for an SWB instrument designed to be sensitive to meaningful change for clients in clinical settings. In small studies comparing clinical and nonclinical samples, the ORS differentiated between them adequately (e.g., *t*-test *p* values of < .0001; Duncan, Sparks, Miller, Bohanske, & Claud, 2006; Miller et al., 2003).

Procedure

Client data were obtained at each session through a computer interface or paper-and-pencil version of the instrument. Data were then tabulated and analyzed using Microsoft Excel.

Data Analysis

Means, standard deviations, and pretreatment-posttreatment Pearson correlations were calculated for the entire sample of 15,398 clients, and for the 13 individual therapists as described in the Participants section above (Table III).

Raw effect sizes were calculated based on the *SD* of pretreatment scores (ES_{pre}), the *SD* of the pre-post difference scores (ES_{diff}), and two reference-group-*SD* effect sizes described below (ES_{refPUB} and ES_{refWEB}). In addition, Dunlap et al.'s (1996) repeated-measures-corrected effect size was calculated (ES_{RMC}) along with two comparison *t*-test-based *ES*s: one derived from an independent-groups *t* test (ES_{ti}), and one derived from a paired-group *t* test without Dunlap's correction (ES_{RMU}). Finally, four severity-adjusted effect sizes (ES_{SA}) using the SD_{pre}, SD_{diff}, and the two versions of SD_{ref} were calculated. These 11 effect size formulas were calculated by the following methods.

ES_{pre} was calculated by averaging the pretreatment (first session) scores, subtracting this average from the average posttreatment (last observation carried forward) score, and then dividing the difference (M_{diff}) by the standard deviation of the pretreatment scores (SD_{pre}).

$$ES_{pre} = (M_{post} - M_{pre})/SD_{pre} = M_{diff}/SD_{pre} \quad (2)$$

ES_{diff} was calculated by averaging the pretreatment scores, subtracting this average from the average posttreatment score, and then dividing the

Table III. Descriptive statistics and Pearson correlations for therapist samples and comparison studies

Sample	Setting	*n*	M_{pre} (*SD*)	M_{post} (*SD*)	SD_{diff}	$r_{pre\text{-}post}$
Web software	nd	15,398	20.6 (9.5[a])	27.4 (10.0)	(9.9)	.48
Therapist A	nd	165	17.3 (9.1)	23.2 (11.5)	(11.5)	.40
Therapist B	nd	180	15.7 (8.8)	22.8 (10.5)	(9.6)	.52
Therapist C	nd	203	20.9 (8.9)	29.3 (8.5)	(8.8)	.49
Therapist D	Private practice	182	17.2 (7.6)	29.1 (9.8)	(9.3)	.45
Therapist E	Private practice	191	15.9 (7.4)	28.0 (9.0)	(9.1)	.39
Therapist F	Integrated health practice	364	16.6 (6.8)	28.6 (8.8)	(9.2)	.33
Therapist G	nd	76	17.2 (9.2)	19.7 (10.0)	(11.3)	.31
Therapist H	nd	88	23.0 (9.4)	25.2 (9.8)	(9.5)	.51
Therapist I	nd	61	24.1 (9.9)	30.8 (8.9)	(9.2)	.52
Therapist J	nd	54	24.8 (9.9)	31.3 (8.6)	(8.4)	.59
Therapist K	nd	56	21.1 (8.8)	32.4 (8.2)	(9.9)	.32
Therapist L	nd	110	15.5 (7.9)	23.9 (9.3)	(9.2)	.44
Therapist M	Private practice	55	17.0 (7.6)	27.8 (8.4)	(9.3)	.33
Anker et al., 2009	Family counseling agency	206[b]	18.1 (7.9)	26.4 (10.0)	nd	nd
		204	18.6 (7.0)	21.7 (8.7)	nd	nd
Hafkenscheid et al., 2010	Outpatient clinic	126	19.3 (8.2)	nd	nd	nd
Miller et al., 2003	Family counseling agency	435	19.6 (8.7)	25.7 (8.7)	nd	nd
Miller et al., 2006	Telephone-based EAP	1244[c]	18.3 (6.8[d])	20.8 (nd)	nd	nd
		1568	18.6 (6.8[d])	22.8 (nd)	nd	nd
		3612	19.0 (6.8[d])	24.4 (nd)	nd	nd
Reese et al., 2009	University counseling center	53[e]	18.6 (7.6)	31.3 (6.6)	nd	nd
		18	22.7 (9.7)	29.5 (7.3)	nd	nd
		51	18.7 (10.4)	29.5 (9.6)	nd	nd
		21	19.6 (6.5)	24.3 (7.5)	nd	nd

Note. Pearson correlations are between pretreatment and posttreatment scores. Therapists A, B, C, G, H, I, J, K, and L were therapists within the web software sample. nd = no data available.
[a]*SD* = 9.466, used in all ES_{refWEB} calculations and ES_{pre} for the web software sample.
[b]Two subsamples: therapists receiving feedback and treatment-as-usual (TAU), respectively.
[c]Three observation phases: baseline, therapist feedback training, and evaluation, respectively.
[d]SD_{ref} from a nonclinical sample in Miller et al. (2003).
[e]Four subsamples: university counseling center: feedback and TAU; graduate training clinic: feedback and TAU, respectively.

difference by the standard deviation of client change scores (SD_{diff}).

$$ES_{diff} = (M_{post} - M_{pre})/SD_{diff} = M_{diff}/SD_{diff} \quad (3)$$

ES_{ref} was calculated according to the following formula:

$$ES_{ref} = (M_{post} - M_{pre})/SD_{ref} = M_{diff}/SD_{ref} \quad (4)$$

Two values for SD_{ref} were derived: the unweighted average pretreatment *SD*s from 11 samples in five studies of the pencil-and-paper version of the ORS, yielding an SD_{refPUB} (from published articles) of 7.85 (Anker, Duncan, & Sparks, 2009; Hafkenscheid et al., 2010; Miller et al., 2003, 2006; Reese et al., 2009); and the SD_{pre} from the current web-based sample of 15,398 clients, yielding an SD_{refWEB} of 9.47 (see Table III).

ES_{RMC} was computed according to Dunlap et al.'s (1996; Equation 3) method, where $ES_{RMC} = d$:

$$d = t_C[2(1-r)/n]^{1/2} \quad (5)$$

and where

$$t_C = M_{diff}/\left(SD_{diff}/n^{1/2}\right) \quad (6)$$

The ES_{RMC} is derived from a *t* test of two correlated (paired) means corrected for the covariation of those means (*r* is the Pearson correlation coefficient for the pretreatment and posttreatment scores). For comparison, an uncorrected ES_{RMU} was calculated from a paired-sample *t* test using the conventional formula for converting a *t* statistic to *d* (i.e., without the Dunlap et al. correction):

$$d = t_I(2/n)^{1/2}, \text{or alternatively, } d = 2t/(df)^{1/2} \quad (7)$$

The ES_{ti} was derived from an independent-groups *t* test following this same conventional formula (i.e., without the correction; see Table IV). The ES_{RMC} was calculated using standard Microsoft Excel formula methods, while the ES_{RMU} and ES_{ti} can be calculated with the Excel's Analysis ToolPak add-in.

Table IV. Effect sizes calculated from *t* tests, with and without correction

Sample	ES_{RMC}	ES_{RMU}	ES_{ti}
Web Software	.69	.97	.69
Therapist A	.56	.72	.56
Therapist B	.72	1.04	.73
Therapist C	.96	1.35	.96
Therapist D	1.35	1.82	1.37
Therapist E	1.46	1.87	1.47
Therapist F	1.52	1.85	1.53

Note. ES_{RMC} = repeated-measures effect size using the Dunlap et al. (1996) correction for paired *t* tests; ES_{RMU} = repeated-measures effect size for paired *t* test, uncorrected; ES_{ti} = effect size calculated from a *t* test formula for *d* designed for independent groups.

ES_{SA} was calculated from four different *SD*s using a simple linear (ordinary least squares) regression equation (G. S. Brown, personal communication, July 15, 2006):

$$y = mx + b \quad (8)$$

or

$$\text{predicted posttreatment score} = \text{slope} \times \text{pretreatment score} + \text{intercept}$$

and from the mean of residual change scores. First, the slope and intercept were acquired from the pretreatment and posttreatment scores in the large web-software-based sample (n = 15,398). ES_{SApre} and ES_{SAdiff} were calculated from the pretreatment and difference-score *SD*s respectively. ES_{ref} was calculated two ways: from the *SD* of this sample (ES_{SAWEB}) and from the average *SD* of published paper-and-pencil samples (ES_{SAPUB}) as described in the ES_{ref} section above (see Table V).

The results of the following calculations were assembled to construct each ES_{SA}.

Reference Group Data Calculations

1. SD_{refPUB} and SD_{refWEB} (as well as sample-based SD_{pre} and SD_{diff})
2. Slope and intercept
3. Mean pretreatment score and mean posttreatment score
4. Mean change score (mean pretreatment score minus mean posttreatment score)

Sample-Based Calculations

5. Predicted client posttreatment scores (using slope and intercept from reference group)
6. Difference between predicted and actual posttreatment scores (residual scores)
7. Mean residual score

$$ES_{SA} = [(\text{reference group's mean change score}) + (\text{sample's mean residual score})] / \text{SD}$$

The ES_{SA} is obtained by adding Step 4 and Step 7, and dividing the sum by one of the *SD*s in Step 1. An alternative method requires the same Steps 1–4 above, and then:

Table V. Comparison of effect size formulas for the web software sample and individual therapists

Sample	ES_{pre}	ES_{diff}	ES_{refPUB}	ES_{refWEB}	ES_{RMC}	ES_{SApre}	ES_{SAdiff}	ES_{SAPUB}	ES_{SAWEB}	Max–Min
Web software	.71	.68	.86	.71	.69	.71	.68	.86	.71	.18
Therapist A	.64	.51	.75	.62	.56	.47	.37	.55	.45	.38
Therapist B	.80	.74	.90	.75	.72	.53	.49	.60	.50	.41
Therapist C	.94	.95	1.07	.88	.96	.95	.97	1.09	.90	.20
Therapist D	1.58	1.29	1.52	1.26	1.35	1.36	1.11	1.31	1.08	.50
Therapist E	1.63	1.32	1.54	1.28	1.46	1.31	1.07	1.25	1.03	.59
Therapist F	1.76	1.31	1.53	1.26	1.52	1.48	1.09	1.28	1.06	.71
Therapist G[a]	.27	.22	.32	.26	.26	.09	.07	.11	.09	.25
Therapist H[a]	.23	.23	.28	.23	.23	.35	.35	.42	.35	.20
Therapist I[b]	.67	.72	.85	.71	.71	.85	.91	1.07	.89	.40
Therapist J[b]	.66	.78	.83	.69	.70	.86	1.01	1.09	.90	.43
Therapist K[c]	1.30	1.14	1.45	1.20	1.34	1.32	1.17	1.48	1.22	.33
Therapist L[c]	1.07	.92	1.07	.89	.97	.75	.64	.75	.62	.45
Therapist M	1.42	1.17	1.38	1.14	1.35	1.18	.97	1.15	.95	.47

Note. Slope = .51; intercept = 16.80 for ES_{SA} calculations; Max–Min = difference between the highest and lowest obtained *ES* among the nine calculation methods. Therapists A, B, and C were randomly drawn from the large (n = 15,398) web software sample among those with $n > 99$. Therapists D, E, and F were independent psychologists with $n > 99$, from a sample-of-convenience. Therapists G, H, I, J, K, and L were drawn from the large (n = 15,398) web software sample among those with $n > 50$.
Therapist M was a US psychologist in private practice.
[a]Therapist chosen from low-effectiveness ($ES_{pre} < .30$) therapists in web software sample.
[b]Therapist chosen from medium-effectiveness ($.60 < ES_{pre} < .70$) therapists in web software sample.
[c]Therapist chosen from high-effectiveness ($ES_{pre} > 1.00$) therapists in web software sample.

Alternative Sample-Based Calculations

5. Predicted client posttreatment scores (using slope and intercept from reference group)
6. Difference between predicted and actual posttreatment scores (residual scores)
7. Residual effect size (residual score / *SD*) for each client
8. Mean residual *ES*

$$ES_{SA} = (\text{sample's mean residual } ES) + (\text{reference group's mean } ES)$$

In this alternative method, the ES_{SA} is obtained by dividing Step 4 by Step 1, and adding the quotient to Step 8. Both methods return the same result; however, the first method requires fewer calculations.

Results

The results of the three *t*-test-based *ES* calculations are given in Table IV. Nine main types of *ES* calculation (four raw, one corrected *t* test, and four severity-adjusted) are given in Table V and Figures 1 and 3. Differences among *ES* calculations for individual therapists were substantial. In the first group of therapists analyzed (Therapists A–F), four of the six therapists obtained a range greater than .40 from among the different *ES* calculations (Table V). The minimum and maximum *ES* values did not consistently result from any particular type of *ES* calculation, except for the three psychologists (D, E, and F) whose *ES* scores were considerably higher than those who came from the large web software sample. These high-performing psychologists all obtained the highest *ES* estimates from ES_{pre}, and the lowest *ES* from ES_{SAWEB} (see Table V and Figure 1).

Figure 2 shows a comparison between ES_{RMC} and four different composite *ES* scores for the web-based sample and Therapists A–F: (1) the average of all four raw *ES* calculations: ES_{pre}, ES_{diff}, ES_{refPUB}, and ES_{refWEB}; (2) the average of all four severity-adjusted *ES* calculations: ES_{SApre}, ES_{SAdiff}, ES_{SAPUB}, and ES_{SAWEB}; (3) a composite of: the average raw *ES* score (i.e., Composite 1), average severity-adjusted *ES* score (i.e., Composite 2), and the corrected repeated-measures *ES* score (i.e., ES_{RMC}); and (4) a composite of the ES_{pre}, ES_{RMC}, and ES_{SApre} scores. While the differences between Composite 1 (average raw *ES*) and Composite 2 (average severity-adjusted *ES*) were substantial at the individual therapist level (often > .20), the differences between Composites 3 and 4, being drawn from across the raw, severity-adjusted, and repeated-measures *t*-test *ES* scores, were low—though greater for the higher-performing psychologists (all < .20). The differences between the ES_{RMC} and Composites 3 and 4 also were low, with the differences only rising as high as .11 (the difference between the ES_{RMC} score and Composite 3 *ES* for Therapist F).

The consistency with which the ES_{RMC} avoided the extreme ends of the range of *ES* formulas was an unexpected finding. To test the replicability of results shown in Figures 1 and 2, outcome data from

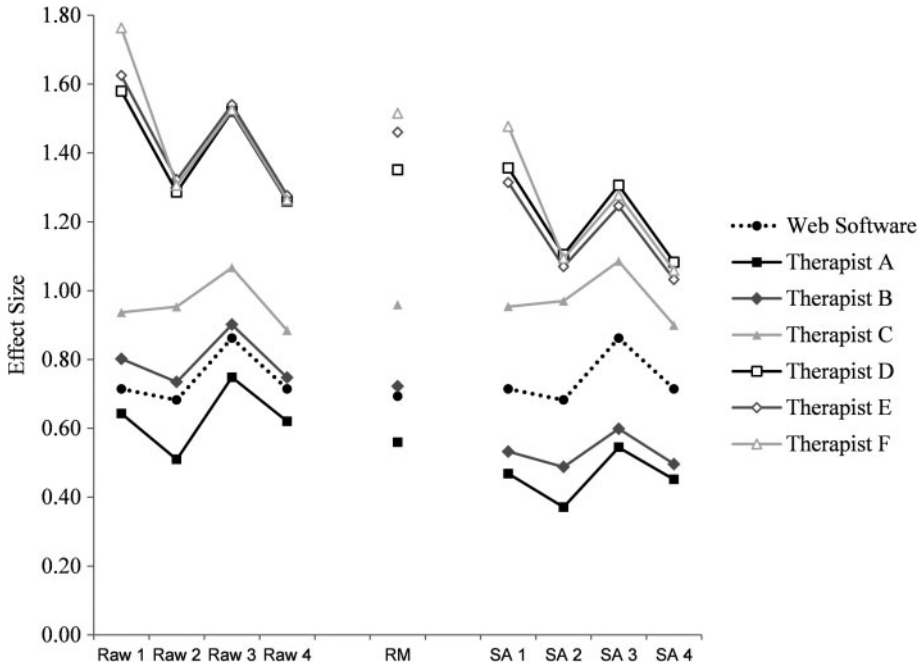

Figure 1. Nine *ES* variations for web software sample and Therapists A–F. Raw 1 = ES_{pre}, Raw 2 = ES_{diff}, Raw 3 = ES_{refPUB}, Raw 4 = ES_{refWEB}, RMC = ES_{RMC}, SA 1 = ES_{SApre}, SA 2 = ES_{SAdiff}, SA 3 = ES_{SAPUB}, SA 4 = ES_{SAWEB}.

seven additional psychotherapists using the ORS were analyzed according to the same methods as described above. Results of the *ES* analysis for these seven therapists (Therapists G–M) are shown in Table V and Figures 3 and 4.

Similar to the results with the first group of therapists shown in Table V, the differences among *ES* calculations for these seven therapists were substantial. Four of the seven therapists obtained a range greater than .40 among the different *ES* calculation methods. Also similar to the results with the first group, the minimum and maximum *ES* values did not consistently result from any particular type of *ES* calculation. In contrast with the first

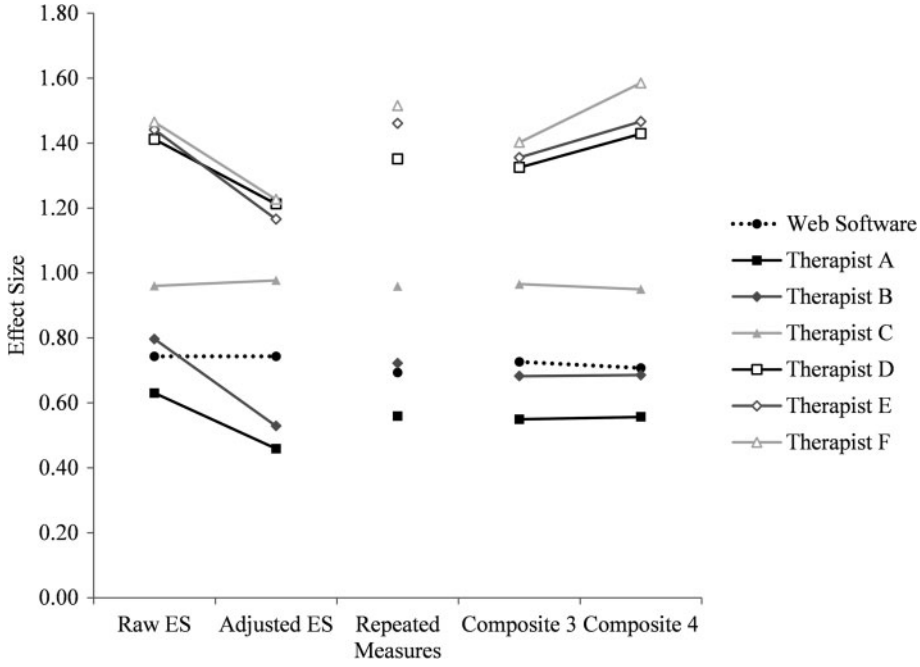

Figure 2. Composite *ES* scores and corrected repeated-measures *ES* for web software sample and Therapists A–F. Raw *ES* = Composite 1: Average of all four raw effect size calculations. Adjusted *ES* = Composite 2: Average of all four severity-adjusted effect size calculations. Repeated-Measures = Dunlap et al.'s (1996) corrected repeated-measures conversion (ES_{RMC}) from *t* statistic. Composite 3 = Average of: average of four raw *ES*, average of four severity-adjusted *ES*, and ES_{RMC}. Composite 4 = Average of: raw *ES* with pretreatment *SD*, severity-adjusted *ES* with pretreatment *SD*, and ES_{RMC}.

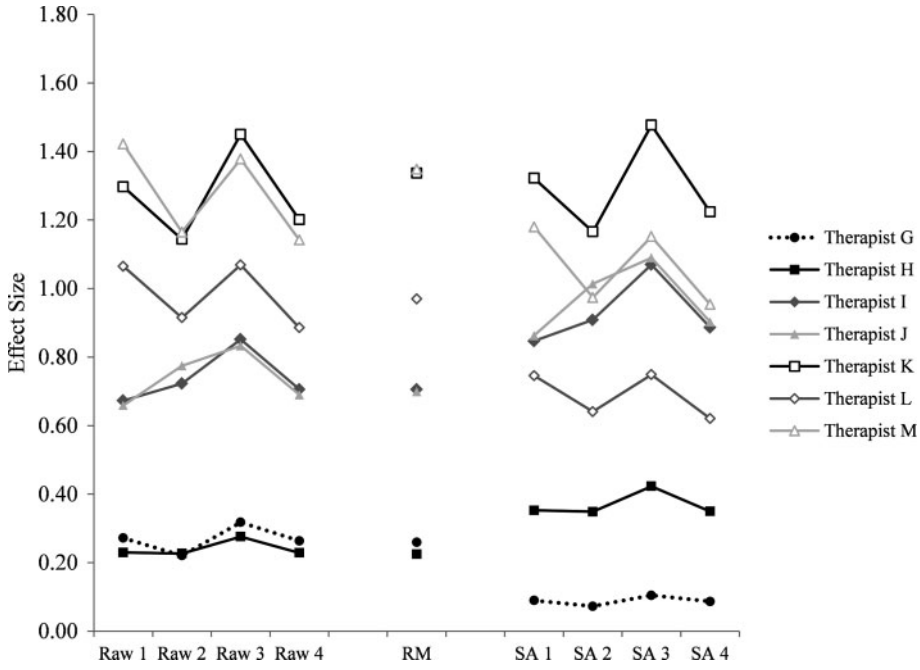

Figure 3. Nine *ES* variations for follow-up sample (Therapists G–M). Raw 1 = ES_{pre}, Raw 2 = ES_{diff}, Raw 3 = ES_{refPUB}, Raw 4 = ES_{refWEB}, RMC = ES_{RMC}, SA 1 = ES_{SApre}, SA 2 = ES_{SAdiff}, SA 3 = ES_{SAPUB}, SA 4 = ES_{SAWEB}.

group, there was not a consistent source of high and low *ES* values from particular formulas for the higher-performing therapists.

Figure 4 shows a comparison between ES_{RMC} and the four different composite *ES* scores for the second group of therapists. As with the first group, the differences between Composite 1 and Composite 2 were substantial at the individual therapist level (often > .20); whereas the differences between Composites 3 and 4 were low (all < .20). The differences between the ES_{RMC} and Composites 3 and 4 were low as well, with the differences only

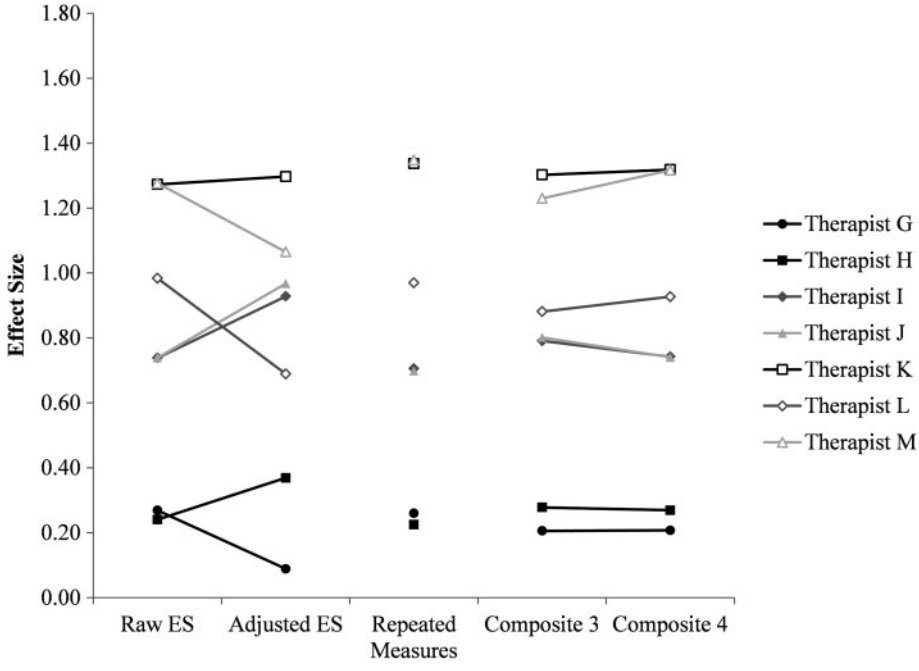

Figure 4. Composite *ES* scores and corrected repeated-measures *ES* for follow-up sample (Therapists G–M). Raw *ES* = Composite 1: Average of all four raw effect size calculations. Adjusted *ES* = Composite 2: Average of all four severity-adjusted effect size calculations. Repeated-Measures = Dunlap et al.'s (1996) corrected repeated-measures conversion (ES_{RMC}) from *t* statistic. Composite 3 = Average of: average of four raw *ES*, average of four severity-adjusted *ES*, and ES_{RMC}. Composite 4 = Average of: raw *ES* with pretreatment *SD*, severity-adjusted *ES* with pretreatment *SD*, and ES_{RMC}.

rising as high as .12 (the difference between the ES_{RMC} score and Composite 3 *ES* for Therapist M).

Discussion

The variety of methods available for measuring *ES* presents the clinician-researcher with a bewildering array of options, any of which may significantly alter the reported effect (some differences between the nine *ES* formulas examined here for individual therapists—see Figures 1 and 3—were as large as the total clinical change commonly attained by clinicians). Three methods of estimating *ES* moderated these extreme values, two of which are practical enough for most clinician-researchers to use as a potentially less biased representation of their *ES* than ES_{pre} or other *ES* calculations that are significantly affected by choice of *SD*.

The first method that provided consistently more moderate results than either the raw or severity-adjusted methods was the Dunlap et al. (1996) corrected conversion of repeated measures *t* to *d* (ES_{RMC}). This method of *ES* calculation can be accomplished using only the sample pretreatment scores, postreatment scores, and a standard computer spreadsheet program (e.g., Microsoft Excel, using paired-two-sample *t* values from its Data Analysis plug-in) and requires little time for calculation.

The second method that moderated the results while remaining practical was Composite 4, the average of ES_{pre}, ES_{SApre}, and ES_{RMC}. This composite *ES* score relied on the SD_{pre} for the raw and severity-adjusted *ES*, and on the SD_{diff} for the ES_{RMC}. Calculating the ES_{SApre} required the calculation of a slope and intercept for a large reference group as well as the pretreatment and posttreatment means of the reference group from which the regression equation was obtained. It did not take into account other (e.g., reference-based) *SD*s that might correct for bias in the sample, but the results were similar to Composite 3, which was based on an averaging of all nine *ES* calculation methods (by clustering them into three *ES* components of raw, severity-adjusted, and repeated-measures *ES*, and then averaging the three components). Composite 3, while taking into account more of the highs and lows of the various *ES* calculation methods, was also the most impractical and time-intensive of the three methods for the interested clinician-researcher to construct, requiring reference-group *SD*s and numerous raw and severity-adjusted *ES* calculations.

With additional published data for appropriate reference groups, some of the variation in *SD* might be minimized for a given sample, yielding a tighter cluster of *ES* scores from more externally valid divisors. Until then, it would appear that clinician-researchers would do well to calculate *ES* values that take into account the vagaries of *SD* selection, case-mix effects, and correlation between pretreatment and posttreatment scores to estimate treatment effectiveness. In the current study, Dunlap et al.'s (1996) ES_{RMC} and Composite 3 and 4 scores (which included ES_{RMC}) significantly mitigated the high variability between *ES* calculations for the large web-based sample and 13 individual therapists tested with these methods.

This study has several limitations affecting the conclusions drawn from it. The data were acquired for only one measure of subjective well-being: the ORS. The generalizability of these results to data from other SWB instruments is unknown, and while correlations between measures of SWB can be moderate or high, response variability between different SWB instruments for the same clients has been demonstrated (e.g., Werbart et al., 2013). There was little or no information available about the various treatment settings, methods, and client demographics from the large web-based sample of therapists. While many researchers have shown that such information has little bearing on outcome (Lambert, 2004), the use of appropriate reference groups (e.g., for standard deviations or the slopes and intercepts of linear regression equations) could provide more rigor in establishing appropriate *ES* divisors and adjustment factors. Similarly, while this does not affect *ES* calculations per se, no information was available for session frequency or frequency of assessment to assess how the "density" of psychotherapy sessions may have affected therapist outcomes, or the extent to which the ORS was actually given session-by-session as clinicians are trained to do.

In this study, only one general approach to clinical change measurement was examined: the SGPP *ES*. Other researchers have shown similar problems with using multiple analytic methods for clinical change in meta-analyses (Ray & Shadish, 1996) and another common approach to clinical change measurement: the *clinically significant change* construct (Bauer, Lambert, & Nielsen, 2004; Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991; Speer & Greenbaum, 1995). These measurement difficulties are likely to be endemic to the enterprise of quantifying change in client-reported well-being, all of which may benefit from "less mathematical wrangling and more empirical testing" (Jacobson et al., 1999, p. 306). Other complex analytical questions have been left unanswered and require further research, including whether—in response to treatment—variability in client scores is different in the beginning, middle, or end of

psychotherapy (which could affect calculations of effect size that take into account variability later in treatment). Providing clients with measures of well-being that are feasible for session-by-session administration and careful coding of sessions that lack the administration of measures might also provide clinician-researchers with more information about drop-out rates and other complex end-of-treatment issues as they relate to the measurement of clinician effectiveness.

Another limitation, although intentional in this study, is the elementary nature of the statistical analyses that were employed. For example, it is well known that the patterns (e.g., slopes and intercepts) of clinical change among individual clinicians vary a great deal, so creating simple severity-adjusted *ES* values for a therapist based on the slope and intercept of a large reference group is merely a way of comparing individual therapists to this norm without explaining why those differences occurred. Yet, researchers who make improvements on the concepts and strategies presented here are advised to do so in a way that remains accessible and open to further critique and refinement by other clinician-researchers wishing to use reliable, valid, and—importantly—feasible methods for analysis and reporting of SGPP effectiveness in real-world clinical practice.

Measurement techniques that are transparent and comprehensible, and that also provide useful and accurate information about clinical outcomes to clinicians, program administrators, and consumers, are difficult to construct. Cohen (1988) cautioned that there was "a certain risk inherent in offering conventional operational definitions" (p. 25) for his "small," "medium," and "large" effect sizes of .2, .5, and .8, respectively, which he posited for independent (not repeated-measure) means. In 1992, he stated that "my intent was that medium ES represent an effect likely to be visible to the naked eye of a careful observer … I set small ES to be noticeably smaller than medium but not so small as to be trivial, and I set large ES to be the same distance above medium as small was below it" and that "the definitions were made subjectively" (1992, p. 156). Leaving aside the issue of how the non-independence of pre versus post measurements should affect this conventional set of *ES* benchmarks, the utility of these benchmarks in estimating therapy effectiveness is questionable if the differences between *ES* calculations for the same client sample often are higher than .4 simply by altering the formula used for measuring a therapist's caseload. The difficulty is compounded by conceptual issues that abound in the operationalization of what "better" means in mental health treatment (Stiles, 2013). The formulas compared in the current study do not address these conceptual issues or the instrumentation concerns (such as ceiling and floor effects, habituation to test items, and sensitivity to change) of how various test items capture client well-being and clinical change over time. Nevertheless, the use of meaningful and accurate change statistics is vital for the growth and accountability of a profession that has been struggling to measure and improve psychotherapy effectiveness for many decades. Refining these statistics from a disorganized array of choices to a more moderate, reliable, and feasible statistic can improve both the accuracy and meaningfulness of outcome measures while clinicians and researchers continue to resolve the deeper methodological challenges still facing the field of psychotherapy outcomes research.

References

Anderson, T., Ogles, B. M., Patterson, C. L., Lambert, M. J., & Vermeersch, D. A. (2009). Therapist effects: Facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology*, *65*, 755–768. doi:10.1002/jclp.20583

Andrews, W., Twigg, E., Minami, T., & Johnson, G. (2011). Piloting a practice research network: A 12-month evaluation of the Human Givens approach in primary care at a general medical practice. *Psychology and Psychotherapy: Theory, Research and Practice*, *84*, 389–405. doi:10.1111/j.2044-8341.2010.02004.x

Anker, M. G., Duncan, B. L., & Sparks, J. A. (2009). Using client feedback to improve couple therapy outcomes: A randomized clinical trial in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *77*, 693–704. doi:10.1037/a0016062

Australian Mental Health Outcomes and Classification Network. (2005). *Adult National Outcomes & Casemix Collection Standard Reports* (1st ed., Version 1.1). Brisbane: Author.

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology*, *77*, 203–211. doi:10.1037/a0015235

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, *75*, 842–852. doi:10.1037/0022-006X.75.6.842

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

Barkham, M., Mellor-Clark, J., Connell, J., & Cahill, J. (2006). A core approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling & Psychotherapy Research*, *6*, 3–15. doi:10.1080/14733140600581218

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*, 60–70. doi:10.1207/s15327752jpa8201_11

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278. doi:10.1111/j.2044-8317.1988.tb00901.x

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruiz, M. (2001). The Pennsylvania Practice Research Network and future possibilities for clinically meaningful and scientifically rigorous psychotherapy effectiveness research. *Clinical Psychology: Science and Practice*, *8*, 155–167. doi:10.1093/clipsy/8.2.155

Bringhurst, D. L., Watson, C. S., Miller, S. D., & Duncan, B. L. (2006). The reliability and validity of the outcome rating scale: A replication study of a brief clinical measure. *Journal of Brief Therapy*, *5*, 23–29.

Brown, G. S., Dreis, S., & Nace, D. K. (1999). What really makes a difference in psychotherapy outcome? Why does managed care want to know? In M. A. Hubble, B. L. Duncan, and S. D. Miller (Eds.), *The heart and soul of change: What works in therapy* (pp. 389–406). Washington DC: American Psychological Association Press.

Brown, G. S. (2006). Accountable Behavioral Health Alliance: Non-Clinical Performance Improvement Project: Oregon Change Index. Retrieved from http://www.clinical-informatics.com/ABHA/OCI%20PIP.doc

Campbell, A., & Hemsley, S. (2009). Outcome rating scale and session rating scale in psychological practice: Clinical utility of ultra-brief measures. *Clinical Psychologist*, *13*, 1–9. doi:10.1080/13284200802676391

Client Voice Innovations. (2013). ASIST outcome management software. Retrieved April 18, 2013, from http://www.clientvoiceinnovations.com/index.html

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

CORE System Trust. (2013). CORE information management systems. Retrieved April 18, 2013, from http://www.coreims.co.uk/

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes in the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*, 245–260. doi:10.1080/02643294.2010.513967

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

Crits-Christoph, P., & Mintz, J. (1991). Implication of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology*, *59*, 20–26. doi:10.1037/0022-006X.59.1.20

Duncan, B. L. (2012). The Partners for Change Outcome Management System (PCOMS): The Heart and Soul of Change Project. *Canadian Psychology*, *53*, 93–104. doi:10.1037/a0027762

Duncan, B., Sparks, J., Miller, S., Bohanske, R., & Claud, D. (2006). Giving youth a voice: A preliminary study of the reliability and validity of a brief outcome measure for children, adolescents, and caretakers. *Journal of Brief Therapy*, *5*, 71–87.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. doi:10.1037/1082-989X.1.2.170

Ed & Psych Associates. (2011). Effect size calculator. Retrieved April 18, 2013, from http://download.cnet.com/Effect-Size-Calculator/3000-2053_4-75335275.html?tag=main;dropDownForm

Eid, M., & Larsen, R. J. (2008). *The science of subjective well-being*. New York: Guilford Press.

FIT-Outcomes ApS. (2013). FIT-Outcomes. Retrieved April 18, 2013, from http://www.fit-outcomes.com/

Hafkenscheid, A., Duncan, B. L., & Miller, S. D. (2010). The Outcome and Session Rating Scales: A cross-cultural examination of the psychometric properties of the Dutch translation. *Journal of Brief Therapy*, *7*, 1–12.

Health Factors. (2013). Myoutcomes. Retrieved April 18, 2013, from http://www.myoutcomes.com/

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects. *Journal of Consulting and Clinical Psychology*, *67*, 300–307. doi:10.1037/0022-006X.67.3.300

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. doi:10.1037/0022-006X.59.1.12

Kim, D. M., 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

Lambert, M. J. (2004). *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.). New York: Wiley.

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., … Burlingame, G. M. (2004). *Administration and scoring manual for the Outcome Questionnaire-45*. Salt Lake City, UT: OQ Measures.

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically derived and rationally derived methods for identifying clients at risk for treatment failure. *Clinical Psychology and Psychotherapy*, *9*, 149–164. doi:10.1002/cpp.333

Leichsenring, F., & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. *Journal of the American Medical Association*, *300*, 1551–1565. doi:10.1001/jama.300.13.1551

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

May, H. (2004). Making statistics more meaningful for policy and research and program evaluation. *American Journal of Program Evaluation*, *25*, 525–540.

McMillen, C. J., Lenze, S. L., Hawley, K. M., & Osborne, V. A. (2009). Revisiting practice-based research networks as a platform for mental health services research. *Administration and Policy in Mental Health and Mental Health Services Research*, *36*, 308–321. doi:10.1007/s10488-009-0222-2

Miller, S. D., & Duncan, B. L. (2004). *The Outcome and Session Rating Scales: Administration and scoring manual*. Chicago, IL: Institute for the Study of Therapeutic Change.

Miller, S. D., Duncan, B. L., Brown, J., Sparks, J. A., & Claud, D. A. (2003). The outcome rating scale: A preliminary study of the reliability, validity, and feasibility of a brief visual analog measure. *Journal of Brief Therapy*, *2*, 91–100.

Miller, S. D., Duncan, B. L., Sorrell, R., Brown, G. S., & Chalk, M. B. (2006). Using outcome to inform therapy practice. *Journal of Brief Therapy*, *5*, 5–22.

Miller, S. D., Hubble, M. A., Chow, D. L., & Seidel, J. A. (2013). The outcome of psychotherapy: Yesterday, today, and tomorrow. *Psychotherapy*, *50*, 88–97. doi:10.1037/a0031097

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386. doi:10.1177/1094428106291059

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs, *Psychological Methods*, *7*, 105–125. doi:10.1037//1082-989X.7.1.105

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielsen, L., Dayton, D. D., & Vermeersch, D. A. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology*, *62*, 1157–1172. doi:10.1002/jclp.20272

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447. doi:10.1037/1082-989X.8.4.434

OQ Measures. (2013). OQ measures: The measure of mental health vital signs. Retrieved April 18, 2013, from http://www.oqmeasures.com

Psytek Ltd. (2005). ClinTools software. Retrieved April 18, 2013, from http://clintools.com

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Counseling and Clinical Psychology*, *64*, 1316–1325. doi:10.1037/0022-006X.64.6.1316

Reese, R. J., Norsworthy, L. A., & Rowlands, S. R. (2009). Does a continuous feedback system improve psychotherapy outcome? *Psychotherapy Theory, Research, Practice, Training*, *46*, 418–431. doi:10.1037/a0017901

Rodriguez, B. F., Weisberg, R. B., Pagano, M. E., Machan, J. T., Culpepper, L., & Keller, M. B. (2004). Frequency and patterns of psychiatric comorbidity in a sample of primary care patients with anxiety disorders? *Comprehensive Psychiatry*, *45*(2), 129–137. doi:10.1016/j.comppsych.2003.09.005

Seidel, J. A., & Miller, S. D. (2012). Manual 4: Documenting change: A primer on measurement, analysis, and reporting. In B. Bertolino, & S. D. Miller (Eds.), *ICCE manuals on feedback-informed treatment* (Vols. 1–6). Chicago: ICCE Press.

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). *Psychological Methods*, *8*, 524–534. doi:10.1037/1082-989X.8.4.524

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*, 1044–1048. doi:10.1037/0022-006X.63.6.1044

Stiles, W. B. (2013). The variables problem and progress in psychotherapy research. *Psychotherapy*, *50*, 33–41. doi:10.1037/a0030569

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*, 473–481. doi:10.1037/0022-0167.51.4.473

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology*, *73*, 914–923. doi:10.1037/0022-006X.73.5.914

Wang, Y., & Duan, N. (in preparation). Relative efficiency of longitudinal, endpoint, and change score analyses in randomized clinical trials. Retrieved April 18, 2013, from http://www.columbia.edu/~yw2016/longi_short2.pdf

Werbart, A., Levin, L., Andersson, H., & Sandell, R. (2013). Everyday evidence: Outcomes of psychotherapies in Swedish public health services. *Psychotherapy*, *50*, 119–130. doi:10.1037/a0031386

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy: The use of early identification of treatment failure and problem-solving strategies in routine practice. *Journal of Counseling Psychology*, *50*, 59–68. doi:10.1037/0022-0167.50.1.59

Wiarda Group. (2013). The therapy outcome management systems (TOMS). Retrieved April 18, 2013, from http://www.thetomsapp.com