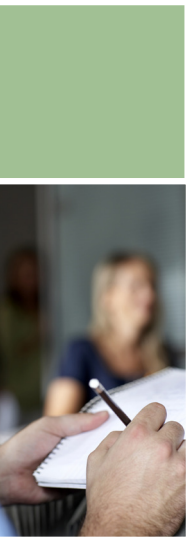


INTERNATIONAL CENTER FOR CLINICAL EXCELLENCE

Manual 4

Documenting Change: A Primer on Measurement, Analysis, and Reporting

ICCE Manuals on Feedback-Informed Treatment (FIT)

INTERNATIONAL CENTER
FOR CLINICAL EXCELLENCE

The ICCE Manuals on Feedback-Informed Treatment (FIT)

Scott D. Miller, Co-Founder, ICCE

Bob Bertolino and Scott D. Miller, Series Editors for ICCE Manuals

The ICCE Manuals on FIT were a collaborative effort. The development team included: Rob Axsen, Susanne Bargmann, Robbie Babbins-Wagner, Bob Bertolino, Cynthia Maeschalck, Scott D. Miller, Bill Robinson, Jason Seidel, and Julie Tilsen.

Manual Authors:

Manual 1: What Works in Therapy: A Primer
Bob Bertolino, Susanne Bargmann, Scott D. Miller

Manual 2: Feedback-Informed Clinical Work: The Basics
Susanne Bargmann, Bill Robinson

Manual 3: Feedback-Informed Supervision
Cynthia Maeschalck, Susanne Bargmann, Scott D. Miller, Bob Bertolino

Manual 4: Documenting Change: A Primer on Measurement, Analysis, and Reporting
Jason Seidel, Scott D. Miller

Manual 5: Feedback-Informed Clinical Work: specific populations and service settings
Julie Tilsen, Cynthia Maeschalck, Jason Seidel, Bill Robinson, Scott D. Miller

Manual 6: Implementing Feedback-Informed Work in Agencies and Systems of Care
Bob Bertolino, Rob Axsen, Cynthia Maeschalck, Scott D. Miller, Robbie Babbins-Wagner

© 2012, International Center for Clinical Excellence

The material in this manual is copyrighted and protected by all laws and statutes, local and international, regarding copyrighted material. No part of this volume may be copied, quoted, or transcribed, in whole or in part, electronically or otherwise, without written permission. The volume and contents are intended for use by the purchaser alone and may not be forwarded, attached, or appended, in whole or in part, electronically or otherwise, without written permission of the publisher.

Introduction to the Series of Manuals

The International Center for Clinical Excellence (ICCE)

The International Center for Clinical Excellence (ICCE) is an international online community designed to support helping professionals, agency directors, researchers, and policy makers improve the quality and outcome of behavioral health service via the use of ongoing consumer feedback and the best available scientific evidence. The ICCE launched in December 2009 and is the fastest growing online community dedicated to excellence in clinical practice. Membership in ICCE is free. To join, go to: www.centerforclinicalexcellence.com.

The ICCE Manuals on Feedback-Informed Treatment (FIT)

The ICCE Manuals on Feedback-Informed Treatment (FIT) consist of a series of six guides covering the most important information for practitioners and agencies implementing FIT as part of routine care. The goal for the series is to provide practitioners with a thorough grounding in the knowledge and skills associated with outstanding clinical performance, also known as the ICCE Core Competencies. ICCE practitioners are proficient in the following four areas:

Competency 1: Research Foundations

Competency 2: Implementation

Competency 3: Measurement and Reporting

Competency 4: Continuous Professional Improvement

The ICCE Manuals on FIT cover the following content areas:

- Manual 1: What Works in Therapy: A Primer
- Manual 2: Feedback-Informed Clinical Work: The Basics
- Manual 3: Feedback-Informed Supervision
- Manual 4: Documenting Change: A Primer on Measurement, Analysis, and Reporting
- Manual 5: Feedback-Informed Clinical Work: Specific Populations and Service Settings
- Manual 6: Implementing Feedback-Informed Work in Agencies and Systems of Care

Feedback-Informed Treatment (FIT) Defined

Feedback-Informed Treatment is a pantheoretical approach for evaluating and improving the quality and effectiveness of behavioral health services. It involves routinely and formally soliciting feedback from consumers regarding the therapeutic alliance and outcome of care and using the resulting information to inform and tailor service delivery. Feedback-Informed Treatment (FIT), as described and detailed in the ICCE manuals, is not only consistent with but also operationalizes the American Psychological Association's (APA) definition of evidence-based practice. To wit, FIT involves "the integration of the best available research…and monitoring of patient progress (and of changes in the patient's circumstances – e.g., job loss, major illness) that may suggest the need to adjust the treatment…(e.g., problems in the therapeutic relationship or in the implementation of the goals of the treatment)" (APA Task Force on Evidence-Based Practice, 2006, pp. 273, 276-277).

Manual 4

Documenting Change: A Primer on Measurement, Analysis, and Reporting

This manual explains how to measure clinical change in psychotherapy, and how to use statistics to understand outcome data. Guidelines are provided for statistical formulas and outcome reporting. The Manual provides examples based on the Session Rating Scale (SRS) and Outcome Rating Scale (ORS), but also pertains to other therapeutic alliance and outcome measures. A short quiz, Frequently Asked Questions (FAQ) and a list of references are also included.

The manual is divided into four sections:

1) establishing a valid baseline (the basics of validity and reliability, and how they pertain to choice of instrument and the administration of outcome measures);

2) graphing client results (guidelines and methods for displaying outcome data for clinical use);

3) understanding clinical significance; and

4) understanding effect size and expected trajectories of change.

1. Establishing a valid baseline: Feasibility, validity, and reliability

Accurately measuring a clinical change requires that clinicians first establish a valid baseline from which change is made. This is true not only for assessing client well-being, but also for assessing clinicians' efforts to improve their effectiveness through deliberate effort and training. To measure how much clients improve from therapy, therapists need to have an accurate sense of how they are doing at the very start of therapy. And to measure how much therapists improve from training and professional development, they need to measure their performance prior to any attempts to improve so a comparison can be made. In order to have confidence that the baselines therapists are measuring for clients are valid (in other words, that baselines actually represent something real and relevant), they need to understand how data from the instrument compare with already-established measures or indicators of real-life well-being or distress. But even before that, a measure must be selected that therapists are actually willing to use with virtually every client in every session. A measure that is accurate in measuring change or performance but that is too costly or awkward to use won't be used consistently, and so from the outset will prevent therapists from establishing a valid baseline. So, the measure must be feasible.

Feasibility

When choosing an outcome measure for tracking client well-being over time, therapists should choose instruments that are feasible for continual use – session after session – and that will not lead to administration or scoring fatigue (either for clinicians or clients). In the real world of clinical practice, the use of long checklists or exhaustive mental health batteries for capturing the fine details of distress quickly becomes unmanageable after an initial administration because of the time required for clients to complete them and for clinicians to score and interpret them. Outcomes need to be monitored quickly and continually for them to be clinically useful. In addition, the inability of therapists to predict clients' unscheduled terminations (e.g., dropouts) necessitates regular administrations of short outcome measures. Giving clients a long measure every 3rd, 5th or 10th session, as some have advised, is likely to create problems in capturing an accurate record of a therapist's outcomes (due to many clients' last available score on the measure not being captured at their actual last session), and in terms of therapists remembering when each client is due for another administration. Studies published by a number of researchers have shown the impact of regular alliance and outcome feedback on treatment effectiveness; so for practical reasons, short (4-, 5- or 10-item) versions of longer "standard" instruments (e.g., Outcome Questionnaire 45 [OQ-45] and Clinical Outcomes in Routine Evaluation [CORE-OM]) have been developed and released in recent years. Field implementation in a variety of settings has also shown that instruments with approximately 30 items lead to very low compliance among clinicians.

Using brief measures promotes consistency in administration which in turn leads to increased validity of the data the clinician accumulates. Put another way, if clinicians' datasets are missing 40% of potential client data, one simply cannot know whether the remaining 60% is an accurate and representative portrayal of the therapist's overall performance, or whether some systematic bias is involved in determining which clients were never administered the outcome measures or which are missing baseline (intake) data. Therefore, the feasibility of the measures – because it is so crucial for consistent administration – should be a major consideration in deciding which instruments best serve a clinician or agency for documenting clinical change.

Validity

For instruments that attempt to measure mental health or a change in client functioning, validity refers to the ability of an instrument to accurately measure what it claims to measure. Below, several aspects of validity are presented that the reader may want to consider for choosing or assessing the quality of an outcome instrument for measuring client distress and functioning. Validity statistics are often reported as correlations (how well two different measures track each other, for example), with coefficients ranging from -1.0 to +1.0. Validity increases the further the coefficient is from zero (either in the positive or negative direction).

Construct validity refers to whether a measure actually captures the construct it is intending to capture; for example, whether a "well-being" measure actually measures well-being, and whether an "alliance" measure is actually measuring the alliance. One of the ways of determining construct validity is through tests of concurrent, discriminant, predictive or criterion-related validity. These specific types of validity address the broader construct validity question ("Does this measure really capture the construct we are interested in?") from different angles. The Outcome Rating Scale (ORS) is reported to be a measure of general subjective well-being, so the construct of subjective well-being needs to be defined, and then the construct validity question is: Does the ORS actually measure a client's subjective sense of well-being?

Note: Judgments about the validity of a measure for a client or group of clients require a faithful and accurate administration of the measure without deviating from the standard way of completing the measure. For example, the Outcome Rating Scale instructs clients to rate well-being over the past week. A client may be strongly affected by the events of today and may complete the form in terms of how the client feels today rather than on an average of the past week. By focusing on one day rather than averaging his or her impressions over the course of a week, the client's scores may show much more variability from one session to the next. If many of a therapist's clients score the measure this way, the statistical conclusion validity of those data is at risk (e.g., the data may not be comparable to norms or other data sets). Substantial deviations, including the wording, item order or other instrument alterations can threaten the validity of an instrument. The ethical and appropriate use of psychometric measures requires that they be administered with regard for these issues.

Concurrent validity is the degree to which a measure yields information/data that is similar to another measure that has already been validated as accurately portraying a construct of interest. One of the most common outcome measures that the ORS has been compared to is the Outcome Questionnaire-45.2 (OQ-45.2), which – like the ORS – differentiates self-reports of clients' general well-being (or distress) into subcategories of internal distress, relationship distress and social functioning. Correlations between the ORS and the OQ-45.2 have ranged between -.53 and -.74. Correlations between the Session Rating Scale (SRS) and other measures of alliance have been between .48 and .63 (see Table 1).

Table 1: Selected Concurrent Validity Data for Outcome and Alliance Measures

Measures	Authors	Validity Coefficient
ORS vs. OQ-45.2	Bringhurst et al., 2006	-.69
ORS vs. OQ-45.2	Campbell & Hemsley, 2009	-.74
ORS vs. OQ-45.2	Miller et al., 2003	-.59
ORS vs. YOQ	Duncan et al., 2006	-.53
CORS vs. YOQ Caretaker	Duncan et al., 2006	-.43
SRS vs. WAI	Campbell & Hemsley, 2009	.63
SRS vs. HAQ-II	Duncan et al., 2003	.48

Note: OQ-45.2 is the Outcome Questionnaire-45.2; YOQ is the Youth Outcome Questionnaire; CORS is the Child Outcome Rating Scale; WAI is the Working Alliance Inventory; HAQ-II is the Helping Alliance Questionnaire-II.

Specificity refers to the ability of an instrument to effectively discriminate between two (presumably separate) conditions along a continuum of experience (e.g., well-being and distress) and to be sensitive to change in a clinical sample. If an instrument that reportedly measures "well-being" is given to a group of people who are known to be in distress and are seeking services (for example, clients at their first session

of therapy), and also is given to a group of people who are generally known not to be in distress (e.g., a "nonclinical" or "community" sample of people), and if the scores of these groups tend to be different, then one would say the measure has specificity. Research on the specificity of the ORS (shown in Table 2 below), as with other outcome instruments, often uses "samples of convenience" such as workers in an office building or counseling center staff. In the case of ORS research discriminating between clinical and nonclinical samples, t-tests or effect-size comparisons are often used. The research shows the ORS to have good concurrent validity with the OQ-45.2 and very good specificity. Similarly, studies show good concurrent validity between the SRS and other therapeutic alliance scales.

Table 2: Specificity Data for Selected Outcome Measures

Measure	Authors	Comparison Method	Statistic
ORS	Miller et al., 2003	t-test between clinical and nonclinical samples	p<.00001
ORS	Miller et al., 2003	Repeated-measures effect size comparison between clinical and nonclinical samples	Clin ES: .70 Non ES: .22
ORS & CORS	Duncan et al., 2006	t-test between clinical and nonclinical samples	p<.0001
OQ-45.2	Lambert et al., 2004	t-test between change slopes of clinical and nonclinical samples	p<.05
OQ-45.2	Lambert et al., 2004	Effect size between change slopes of clinical and nonclinical samples	d=.50
OQ-45.2	Lambert et al., 2004	ANOVA between multiple clinical and nonclinical samples	p<.001

Criterion-related validity is concerned with whether an instrument measures something that has a future or real-world impact. An example of criterion-related validity for an alliance measure is the extent to which an alliance score at the outset of therapy predicts the eventual outcome of therapy. For an

outcome scale, an example of criterion-related validity is the extent to which a change in well-being on the ORS given to people in couples therapy is related to the separation and divorce rate of those couples six months later. Table 3 below gives some examples of criterion-related validity statistics.

Table 3: Criterion-Related Validity Data for Selected Outcome and Alliance Measures

Measure	Authors	Comparison Method	Statistic
ORS	Anker et al., 2009	Separation or divorce at 6-month follow-up, with FIT vs. without FIT.	$X^2 = 4.83$ $p<.02$
SRS	Duncan et al., 2003	Correlation between 2nd or 3rd session SRS score and final ORS score	.29
Penn scales (alliance)	Martin, Garske, & Davis, 2000	Various outcome measures in meta-analysis	.29
Working Alliance Inventory	Martin, Garske, & Davis, 2000	Various outcome measures in meta-analysis	.24

As noted above, the manner in which a clinician introduces a measure, the timing of when it is introduced and other factors influence the validity of the score and affect whether the clinician is establishing a valid baseline from which to measure change. Here are a few client factors which the clinician should keep in mind:

- vision problems, reading comprehension, literacy or language difficulties;
- cognitive impairments or neurological deficits;
- distraction by stressors in or outside of the room;
- frustration with the form as "paperwork," or discomfort with computer administration;
- attention not focused on the personal meaning of the questions;

- preoccupation or worry about the clinician's judgments toward the client's answers;
- perception that the clinician is not genuinely interested in the responses;
- negative feelings toward therapist, agency, or referrer;
- rushed administration of the measure or client feels time-pressure

Clinicians should take great care to ensure that outcome and alliance measures are presented so that clients understand the measures, are able to orient to them with their full attention and are encouraged by the clinician's own genuine attitude to answer them as authentically and truthfully as possible, and without any biasing toward a more distressed or more optimal feeling than is accurate for the past week or the time since last measurement (based on the standardized instructions of the measure).

Note: Some variation from total uniformity of administration is unavoidable in the real world, and is even desirable if doing so will make the measures more clinically and personally relevant to the client and therefore more useful and valid in the measurement of clinical change. One example of this is when clinicians administering the ORS for the first time to clients who have come to treatment because of distress with their partner, suggest that the clients focus on their relationship with their partner on the ORS item that pertains to "family, close relationships" even though the item could also include other close relationships in addition to the partner. Another example of a deviation from uniform administration being potentially helpful is delaying the administration of the baseline (first session) measure until the client and clinician have had some time in the session to develop some rapport, in order to maximize the likelihood that the client will feel comfortable or in touch with his or her true level of distress/well-being and will be able and willing to express it accurately on the measure.

Reliability

There are three types of reliability related to the measurement of outcome and alliance: interrater reliability, test-retest reliability and internal consistency reliability. As with many measures of validity, measurements of reliability are usually reported as correlation coefficients, ranging from -1.0 to +1.0.

Interrater reliability refers to the strength of association between the scores on a measure made by one person versus another person. Psychotherapy researchers have long known that the quality of the therapeutic process and client change are in "the eye of the beholder" (see Orlinsky, Rønnestad, & Willutzki, 2004, for a review). A therapeutic alliance is a complicated process between people and could be judged differently by different raters. High interrater reliability is not always important. For example, a particular rater's judgment might be considered more important or valid; or there may be particular reasons why the agreement between raters is low. A weak correlation between the ratings of a child and his or her family member about the child's well-being (as shown in Table 4 below) is likely to be affected by the different information that each person has in making the judgment about the child's well-being, or different motivations to amplify or minimize the portrayal of distress. On the other hand, independent judges watching a video recording of a therapy session will receive identical information and would be expected to have greater agreement. There are a variety of statistics used for assessing interrater reliability, including the Pearson *r* correlation, the Kappa statistic and various types of intraclass correlation coefficients.

Table 4: Interrater Reliability Data for Selected Outcome and Alliance Measures

Measure	Authors	Raters	Statistic
ORS-Adolescent (13-17yo)	Duncan et al., 2006	child vs. caretaker	.45
CORS-Child (<12yo)	Duncan et al., 2006	child vs. caretaker	.63
Penn alliance scales	Martin et al., 2000	Mixed: clients, therapists, observers	.68
WAI (alliance scale)	Martin et al., 2000	Mixed: clients, therapists, observers	.92

Test-retest reliability refers to the consistency with which people answer the same questions on separate occasions. In the case of instruments (such as the ORS) that are designed to be sensitive to clinical change, the test-retest reliability might be expected to be moderately high for people in the community that are tested on separate occasions one or two weeks apart. Likewise, since clinical distress can remain somewhat stable unless interventions occur, the test-retest correlation should be moderately high for a clinical group that is not undergoing any kind of treatment or intervention. Yet, one would expect the test-retest correlation to be low for a clinical group that is in treatment, assuming the treatment has some effect on the client's well-being. In other words, test-retest reliability is a measure of how stable a variable is across time. Severe clinical distress that has been going on for some time is likely to be more stable over the span of a few weeks if no interventions are made than it would be if a person were having effective treatment. So, an outcome measure would be expected to show lower test-retest reliability for people in treatment than for those who are not in treatment or for those who are not distressed. Research (see Table 5) shows moderate to good test-retest reliability for the ORS and CORS in nonclinical samples (similar to the OQ.45.2), with reliability increasing with client age (children show more variability than adults). These measures of well-being are sensitive to changes in a person's day-to-day life – unlike those measuring stable constructs such as intelligence – and therefore, are even expected to yield variable data when rated by people in the general community. The SRS shows good test-retest reliability that is comparable with other alliance measures.

Table 5: Test-Retest Reliability Data for Selected Outcome and Alliance Measures

Measure	Sample	Authors	Time Span	Statistic
CORS	nonclinical child	Duncan et al., 2006	10-21 days	.60
CORS	nonclinical child (caretaker)	Duncan et al., 2006	10-21 days	.51
ORS	nonclinical adolescent	Duncan et al., 2006	10-21 days	.78
ORS	nonclinical adolescent (caretaker)	Duncan et al., 2006	10-21 days	.72
ORS	nonclinical adult	Miller et al., 2003	Time 1 to 3	.58
OQ-45.2	nonclinical adult	Miller et al., 2003	Time 1 to 3	.75
ORS	nonclinical adult	Bringhurst et al., 2006	1-2 weeks	.80
SRS	clinical adult	Duncan et al., 2003	Time 1 to 2	.70
HAQ-II	clinical adult	Duncan et al., 2003	Time 1 to 2	.75
Penn	Not reported	Martin et al., 2000	Not reported	.55
WAI	Not reported	Martin et al., 2000	Not reported	.73

Note: Penn is the Penn Alliance Scales; WAI is the Working Alliance Inventory.

Internal consistency reliability, measured using Cronbach's coefficient α (alpha) with a range from 0.00 to 1.00, is the degree to which different items on a measure are scored similarly by a client (see Table 6). If a measure has several questions that ask about different aspects of a single construct (e.g., "well-being"), then they would be expected to be highly correlated. Cronbach's α for such measures will be high. If a measure has several unrelated items or is designed in a way that increases erratic or inconsistent responses, then Cronbach's α will be lower.

Table 6: Internal-Consistency Reliability Data for Selected Outcome and Alliance Measures

Measure	Sample	Authors	Time of Administration	Statistic
ORS	Nonclinical adult	Bringhurst et al., 2006	1st administration	.91
ORS	Nonclinical adult	Bringhurst et al., 2006	2nd administration	.93
ORS	Nonclinical adult	Miller et al., 2003	1st administration	.87
ORS	Nonclinical adult	Miller et al., 2003	3rd administration	.96
ORS	Clinical adult	Campbell & Hemsley, 2009	Not described	.90
ORS	Not described	Duncan et al., 2006	Not described	.93
CORS	Not described	Duncan et al., 2006	Not described	.84
OQ-45.2	Nonclinical adult	Bringhurst et al., 2006	All administrations	.98
OQ-45.2	Nonclinical adult	Lambert et al., 2004	Not described	.93
OQ-45.2	Clinical adult	Campbell & Hemsley, 2009	Not described	.95
OQ-45.2	Clinical adult	Lambert et al., 2004	Not described	.93
SRS	Clinical adult	Campbell & Hemsley, 2009	Not described	.93
SRS	Clinical adult	Duncan et al., 2003	All administrations	.88
HAQ-II	Clinical adult	Duncan et al., 2003	All administrations	.90
WAI	Clinical adult	Campbell & Hemsley, 2009	Not described	.91

Summary and Recommendations

The first step in establishing a valid baseline from which a client might progress is to make sure that the measures are feasible – that they will actually be provided to every possible client in every possible session so that the information they provide can be used to improve care. After that essential step, clinicians need to guard against using measures that lack any formal research on their validity and reliability. Measures of outcome and alliance ought to be accompanied by a description of the extent to which the data they yield coincides with that attained using established methods for judging/quantifying the same constructs. A poor therapeutic alliance and a lack of early positive change in well-being in the first few sessions of therapy have both been shown to predict a poor therapy outcome and this is one of the chief reasons for tracking the quality of well-being and alliance. A good alliance instrument will have data showing some predictive validity for future clinical outcomes as well as a strong correlation with similarly constructed alliance instruments. Outcome instruments ought to correlate well with previously validated measures that purport to assess the same constructs; and the validity of an outcome instrument is strengthened by its correlation with other criteria of interest (e.g., separation rate for couples therapy; weight loss for weight-reduction therapy). Instruments are expected to be reliable, in that they perform consistently from situation to situation. Yet, when using instruments that are sensitive to change, it is expected that reliability across time will be less than one would be expected to find in an instrument designed to measure stable personality traits or among people not engaged in therapy to change their well-being. Clinicians and administrators who understand the different kinds of validity and reliability – why and when they are important – can critically evaluate the best choices for their clinical setting. It is worth noting that a large review of outcome studies found that in the 1980s, there were over 1,400 outcome measures used for therapy studies, but 840 of them had been used only once and many lacked any standardization. Even one of the most common measures, the Hamilton Rating Scale for Depression, was found to have more than a dozen versions being used. Without standards, validity testing and reliability testing, an outcome instrument will be unable to be compared with benchmarks or across clinicians.

2. Graphing client results: Guidelines and methods for displaying outcome data for clinical use

Graphing well-being and alliance data over time can help make data more useful and easier to understand for both clients and therapists. In this section, information provided in Manual 3 of this series will be reviewed in a discussion of the elements and proper use of graphs.

One of the most useful aspects of graphing client data is the ability to see dramatic clinical changes, a lack of change, small blips in the alliance, cycling of mood, mismatches between client report on a measure and client verbal report in the session and other patterns that would not be visible simply by asking clients or by looking at numbers in a spreadsheet.

Different intake scores are likely to lead to different amounts of change. Clients with very low well-being scores are likely to show the most change, and clients whose well-being is high when they start therapy are likely to show the least amount of change. Another interpretive tool sometimes provided on graphs by outcomes management software is the expected trajectory of change (ETC) from the first session to the sixth or seventh session (sometimes called the benchmark line). The expected trajectory is based on the average amount of change that clients who entered treatment with a similar level of distress would likely experience over the course of several sessions. In the case of a benchmark score (or target score), an expected score at termination is given, based on the client's level of distress at intake and the amount of change a large number of therapy clients have experienced when they have started treatment at various levels of distress. Researchers have found that providing clinicians with feedback about clients' changes relative to an expected amount can improve outcomes for clients. Clinicians can use such signaling methods as a training tool and to improve their responsiveness to a lack of clinical change.

While these expected trajectory curves can provide general or statistical expectations, it is important to remember that they are averages and do not reflect the variability between individual clients. Figure 1 illustrates individuals presenting for therapy with relatively low well-being (Fig. 1a) and high well-being (Fig 1b). Figure 2 illustrates the same concept,

but dividing the group between successful (Fig. 2a) and unsuccessful (Fig. 2b) cases. The dotted line represents the average trajectory while the different colored lines are actual individual client scores. As can be seen in the figures, individual client trajectories can vary from the average from session to session, and this variability occurs for clients starting with especially low well-being and those with high well-being, and whether clients are ultimately successful in therapy or not.

Figure 1. Individual (1a, 1b) and averaged (1c, 1d) trajectories of change in two samples of 20 clients in their first 7 sessions of therapy.

Figure 1a. Below-average ORS scores at intake

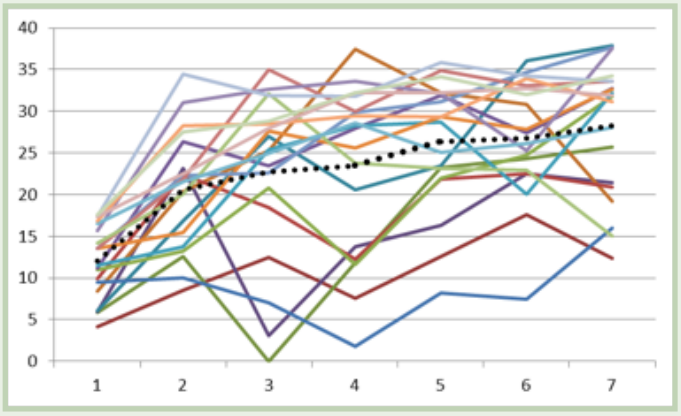


Figure 1b. Above-average ORS scores at intake

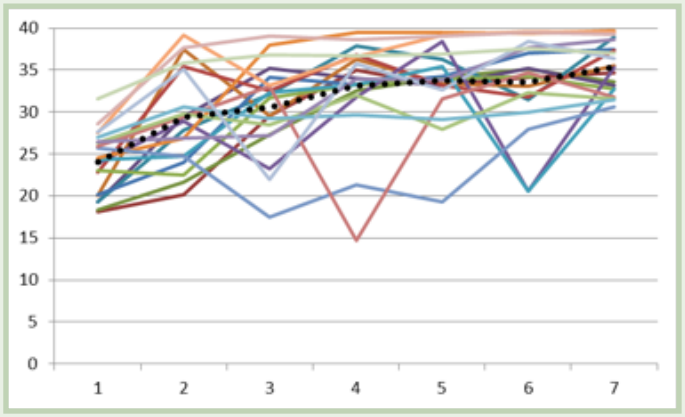


Figure 1c. Average of ORS scores from Fig. 1a

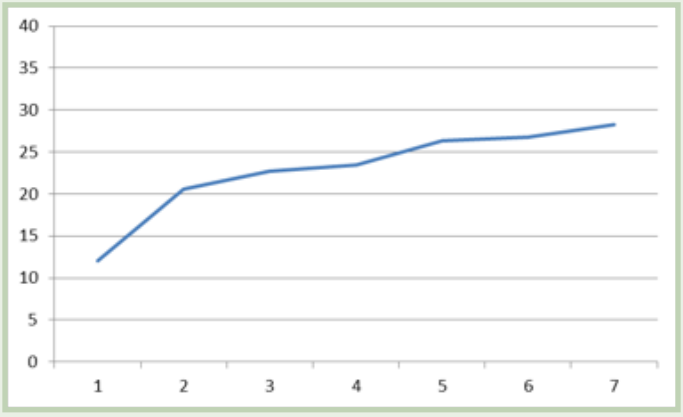


Figure 1d. Average of ORS scores from Fig. 1b

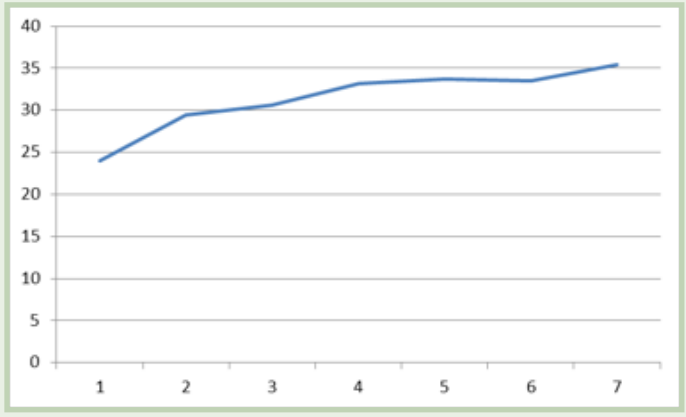


Figure 2. Cases from Figure 1 who started below the ORS clinical cutoff of 25 and showed clinically significant change by termination (2a; N=27) or did not show clinically significant change (2b; N=4)

Figure 2a. Clients with significant improvement

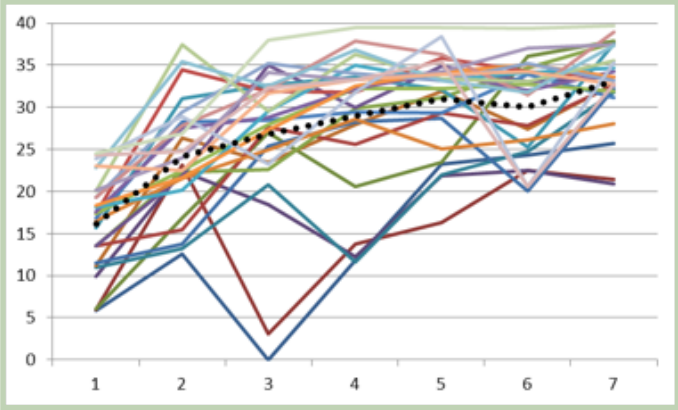
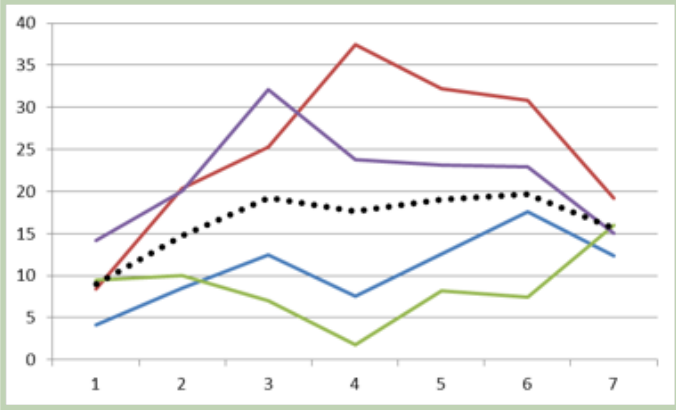


Figure 2b. Clients without significant improvement



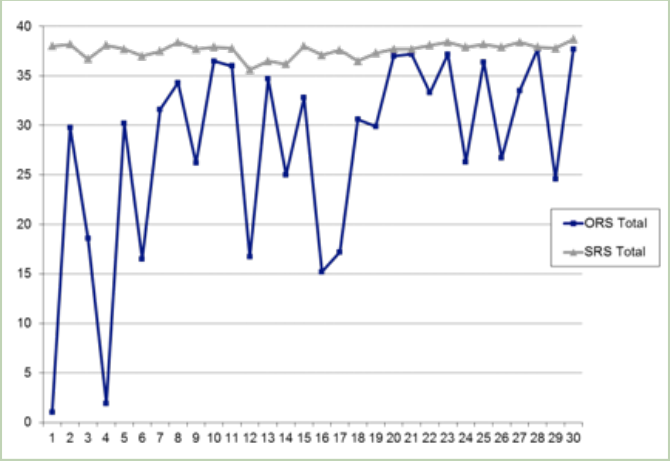
Note to Figure 2: Clinically significant change was defined by an intake score below 25, an improvement by 5 points on the ORS, and a termination score at or above 25.

Significant deviations from the expected trajectory for a given client provide an opportunity for discussion and exploration. At the same time, given the variability among clients, and from session to session, ETCs cannot be used in isolation to determine that a treatment for a particular client is effective or not.

Graphing and interpreting trajectories

Various tools and methods have been developed to help clients and therapists interpret the meaning of trajectories. One of the simplest statistical tools for interpreting a score on a graph is the clinical cutoff (which will be discussed in more detail in Section 3). Comparing a person's score at intake to the clinical cutoff indicates whether he or she is scoring more like those who have started treatment (the "clinical range," below 25 on the ORS) or like those who are not in treatment (the "nonclinical range," 25 or above on the ORS; see Figure 4).

Figure 4. Individual trajectory on the ORS showing the clinical cutoff score (25) for adults.



It is important not to communicate to clients that the clinical cutoff differentiates between being ill and being well. As will be explained in more detail below, the clinical cutoff is actually a midpoint between the well-being ratings given by groups of clients in their first session and groups of nonclients. As such, the clinical cutoff is best used as a rough estimate for either the clinician or both the clinician and the client to inform conversations about the course of treatment and the client's sense of well-being. For example, it is likely that clients who start treatment with scores above the clinical cutoff will show declines in their well-being over the course of treatment, with the greatest risk being associated with the highest intake scores. Clinicians may want their initially high-scoring clients to be aware of this and discuss what that means for their treatment.

Clinicians can use the clinical cutoff and benchmark scores to generate a count of how many clients start therapy below the clinical cutoff and end therapy above the cutoff; or what percentage of clients reach or exceed the benchmark (expected) score by the end of therapy. The next sections provide more details about how to do this.

Summary and Recommendations

Graphs of alliance and outcome data can greatly facilitate discussions between clients and therapists about the process and progress of treatment. These tools can amplify and clarify subtle patterns of change and improve clinicians' vision and hearing about what is happening in the lives of their clients that might otherwise be missed. The more that clinicians know about the construction of these tools, the better able they will be to provide appropriate ways of understanding what they mean. To that point, therapists should be familiar with the signaling or alert systems provided in outcomes-management software, and with any other reference points used in the graphing of outcome data that are meant to enhance the results that clients receive.

3. Understanding clinical significance: The clinical cutoff, reliable change and other concepts

In outcomes measurement and reporting, different statistics can show different things about a clinician's effectiveness. Statistics can focus on how many people have changed in a meaningful way, and other statistics can show how big the changes are that a clinician (or clients) are experiencing. A statistic that is commonly used to show the number of people who have changed following treatment is called "clinically significant change."

Clinically Significant Change: Neil Jacobson and his colleagues (e.g., Jacobson & Truax, 1991) designed various formulas to determine a score for any given clinical measure or instrument that would serve as a dividing line between people who are in the "clinical" range and those who are in the "normal" or "community" range. Jacobson and colleagues argued that although the *effect size* statistic (discussed in Section 4) does show how much of a difference there is between a clinical and nonclinical group, or between a group of clients prior to therapy and after therapy, the effect size does not provide perspective on how clinically meaningful or important that difference is (for an effect size, the amount of variation between people's scores in a group of clients can have a big impact on how big the effect size is). Jacobson et al. argued that large effect sizes could be found that were clinically trivial (if the groups had very low variability between their scores), and that effect sizes don't provide a good way of knowing when a treatment should be considered "effective" from a common sense (rather than a statistical) perspective. Instead, they defined effective through the concept of "clinically significant change," which they measured in several ways. Their basic idea was that for treatment to be considered "clinically significant," the client should experience an amount of change that is beyond a trivial amount of "day-to-day" fluctuation, and the client's distress should change from being within a "clinically distressed" range of scores to being within a typical "community" range of scores. For example, if a weight-loss treatment led to consistent but small changes (e.g., 5% of body weight) in weight for a group of obese patients, this might lead to a large effect size if the variability between patients was small. However, it would not be a clinically significant change for these patients.

Another way of describing these elements is to say that a nontrivial amount of positive clinical change is an "improvement" or a "reliable change" (e.g., 5% of weight); a nontrivial amount of negative clinical change is a "deterioration" or "reliable worsening"; and an "improvement" from a score in the clinical range of scores to a score in the nonclinical range of scores could be considered a "recovery" because it represents a person's change that makes his or her score indistinguishable from a person in the community or nontreatment-seeking population (e.g., moving from within the "obese" range to the "overweight" range of weight, body mass index, etc.). Finally, an amount of change that is within a range that might be expected from normal variation or "error" in scores is called "unchanged" or "uncertain," and scores falling close to the clinical cutoff, whether before or after treatment (taking into account measurement error) would also be called "uncertain" or "unclassifiable" (though this last example is not commonly used).

The clinical cutoff: The clinical cutoff is a statistic designed to help researchers and clinicians think about what separates seriously or "clinically" distressed people (that is, those who presumably want or would benefit from professional help of some kind) from people who are within the usual range of well-being. Although there are several kinds of clinical cutoff, the one most frequently used is "criterion c" which requires knowing the level of distress in a "functional" part of the population and the level of distress in a "dysfunctional" part of the population. The cutoff is essentially a midpoint between the average of each of these group's scores. So a score that is on one side of the cutoff would be more likely to come from someone in the dysfunctional group, and a score on the other side is more likely to come from someone in the functional group. The cutoff is the line that best separates scores between these groups.

Note: For the ORS, previous authors have estimated the clinical cutoff for adults to be 25. Total ORS scores above 25 resemble the scores that nonclinical samples of adults are likely to have, and scores below 25 more closely resemble the scores that people have when they come for their first therapy session. For adolescents aged 13-17, the clinical cutoff has been set to 28, and for children 6-12 years old completing the CORS, the clinical cutoff is 32.

Reliable change: To know whether the amount of change that a client experiences is beyond the "normal" fluctuations of well-being that anyone might experience from week to week, the Reliable Change Index (RCI) was devised. The RCI estimates the amount of change (number of points) on a measure beyond which one can be confident that the change is more than just measurement error (e.g., chance, maturation, and noise). A client whose change is "reliable," based on this number of points, is considered to be "improved" or "reliably improved," while a client whose score declines by at least the RCI is categorized as "reliably worse." This improvement is not the same as "clinically significant change" because (as discussed above) clinically significant change requires not just a statistically reliable amount of change, but also change from a clinical level of functioning to a nonclinical level of functioning (crossing over the clinical cutoff).

Note: For the ORS, previous authors have estimated a 5-point change to be a reliable change. Clinically significant change would be a score that changed from below the cutoff to a score that is equal to or greater than the cutoff, and also changed by at least 5 points. For example, an intake score of 19, and a termination score of 26 would represent clinically significant change. The intake score is below the clinical cutoff of 25, the termination score is at least 25 and the difference between the scores is at least 5 points. An intake score of 26 and a termination score of 37 would not represent clinically significant change because although the change was considerably more than 5 points (i.e., reliably improved), the client started in the nonclinical range and finished in the nonclinical range.

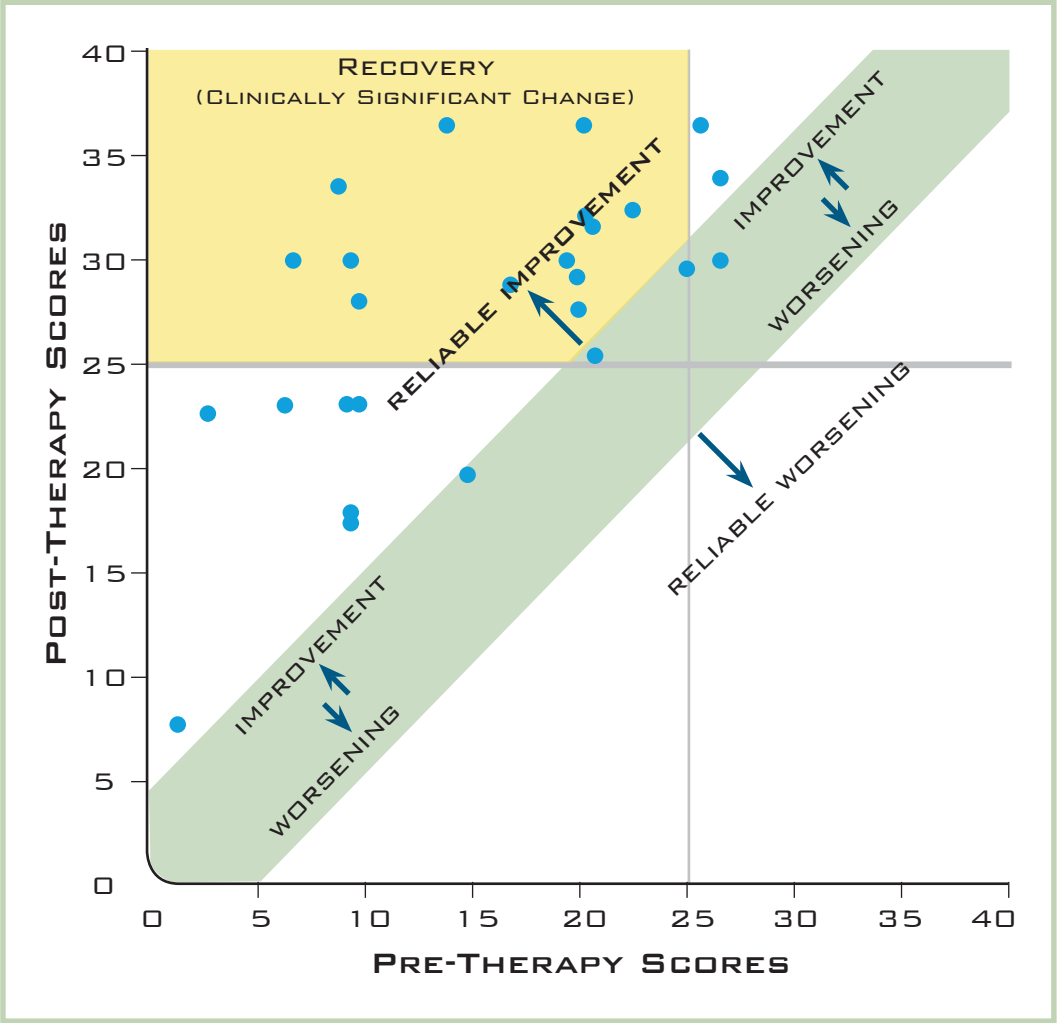
Statistically, "reliable change" is the amount of change (e.g., between pre-therapy and post-therapy scores), divided by the amount of "spread" in the scores that would be expected if no actual change occurred (this spread is called the "standard error of the difference" or S_{diff}; Jacobson & Truax, 1991). If there were literally no spread or fluctuation in scores when no change occurred, one could assume that any difference in scores between the beginning and end of treatment was statistically "real" clinical change. But people vary from time to time in how they feel, even if not in a way that is clinically important. So therapists need

to know how much change is meaningful change on a measure. The RCI is that "big enough change to pay attention to."

The amount of fluctuation likely to be seen in a large group of people who are not in any kind of therapy or "change process" can show how much fluctuation is "normal" for the instrument, and therefore how much more change should be required to be considered important and meaningful. This is where the "standard error of the difference" or S_{diff} comes in. The S_{diff} is made up of a few things, but the two main elements are the standard deviation and the test-retest reliability of the measure (Jacobson & Truax, 1991). The bigger the standard deviation (bigger means more error) and the smaller the reliability of the measure (smaller means more error), the bigger change should be required of a measure to be considered "real" change or "reliable" change. Test-retest reliability is used because it is a measure of how much fluctuation in well-being is likely due to variations just from time passing, without any purposeful attempt to create clinical change. Returning to the weight-loss example, if there is a lot of variation in weight between people at the start of a weight-loss study (i.e., a large standard deviation), and if it is common for people who are obese to fluctuate in weight by 3-5% from week to week (low test-retest reliability), we would need a larger amount of weight change to serve as a "reliable" change beyond the individual differences and the individual, week-to-week levels of change commonly seen. On the other hand, if the range of weight within the "obese" range is quite narrow (small standard deviation), and if it is common to not fluctuate more than 1-2% from week to week (high test-retest reliability), we would not need as large an amount of weight to signal that a reliable change had occurred.

Figure 8 shows an example of a scatterplot (sometimes called a "Jacobson Plot") which can be used to show which clients experience reliable change, clinically significant change, no change, unclassifiable change or worsening. By selecting any pre-therapy score from the horizontal axis and a post-therapy score on the vertical axis and then locating the intersection between the two scores inside the graph, one can determine how many clients would be considered recovered, reliably improved or worsened, etc., based on the clinical cutoff and RCI.

Figure 8. Scatterplot of ORS scores, showing categories of change (adapted from Jacobson & Truax, 1991)



If a clinician wants to tally his or her clients' change without the use of this kind of graph, it is advisable to first separate those clients who started therapy in the clinical range from those who started therapy already above the clinical cutoff. The clinician can first tally how many "clinical" clients showed reliable improvement or worsening, and how many "nonclinical" clients showed reliable improvement and worsening. The clinician can count how many clients in each group showed no (reliable) change, and finally how many clients who started in the clinical range showed clinically significant change (or "recovery").

4. Understanding effect size: Methods for calculation and the use of benchmarking and expected trajectories of change

A statistic that is often used to measure the amount of change (the size of change overall, rather than how many people changed) is called the effect size (ES). An effect size is a way of expressing how much of a difference has occurred between two groups of people (or a group of people at the start of therapy and that same group at the end of therapy) compared with the variation that normally occurs from person to person. For example, consider a study conducted to determine the effect of eating more vegetables on the number of common colds each participant has over five years. At the start of the study, the average participant had five colds in the previous five years (averaging one cold per year), and at the end of the five-year vegetable-eating study, the average person only had 3.5 colds. On the one hand, if virtually every person in the study had five colds before the study and three or four colds after the study, this consistent drop of 20-40% would be very impressive, given that virtually everyone could be assured of fewer colds. But on the other hand, if the number of colds the participants had in the previous five years averaged five colds, but ranged widely (between 0 and 16 colds over five years), and if some people had more colds at the end of the study (but the whole group still averaged 3.5 at the end), the effect would be less impressive even though the average amount of change was the same for the whole group. The effect would be less "big" in terms of importance because the difference occurred with a lot of variation and individual differences, and there would be little confidence about how eating more vegetables would affect any particular person.

The effect size for a group's change in well-being is considered a "within-group" or "repeated-measures" difference because it is a comparison of a group with itself rather than a comparison between two different groups. This kind of effect size is a measurement of the average change in well-being (from the start of therapy to the end of therapy), taking into account the amount of variation in the group members' well-being at the beginning. By taking into account how much variation there is among different people who start therapy, the average amount of change that a

clinician's clients show can be expressed in terms of whether this is a relatively big difference or partly a matter of a naturally big range and fluctuation of well-being among people to begin with. Calculating the "raw" effect size merely requires subtracting one score from the other (the average termination score "m_2" minus the average intake score "m_1") and dividing that by the standard deviation of all of the intake scores). The standard deviation "standardizes" the change by expressing it in terms of the amount of variation that is typically seen among people using that measure, before any treatment intervention has occurred (an intervention like therapy will typically increase the variation by causing scores to spread: some people will respond a lot and some not at all to an intervention).

$$Raw\ ES = \frac{(m2\text{-}m1)}{SDintake}$$

In this section, methods for calculating raw effect sizes in Microsoft Excel will be described. Excel formulas are entered in an empty cell starting with an equals sign followed by the appropriate formula or function to be calculated. The "Formulas" tab at the top of the Excel toolbar provides a wide range of formulas and the statistical formulas are listed under "More Functions." Below are instructions for calculating the means, standard deviations and raw effect size for a sample of ORS data.

Data should be arranged so that clients are in rows, and client data are in columns (e.g., Client ID in column A, Intake Scores in column B, Termination Scores in column C):

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3

To calculate the mean of the Intake Scores in Column B, for Clients 1-3, the following formula is typed into an empty cell: =average(B3:B5) The location of the cell used for the formula is unimportant except for the user's convenience, because the formula contains the information for which cells have the data to be calculated. For this example, the mean will be calculated in Cell B6 (column B, row 6):

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		**= average (B3:B5)**	

When the return key is pressed after entering this formula, the cell shows the result:

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		**15.43333**	

To calculate the mean of the Termination Scores in Column C, for Clients 1-3, the following formula is typed into Cell C6 (column C, row 6): =average(C3:C5)

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		15.43333	**= average (C3:C5)**

When the return key is pressed after entering this formula, the cell shows the result:

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		15.43333	**30.9**

To calculate the standard deviation (SD) of the Intake Scores in Column B, for Clients 1-3, the following formula is typed into Cell B7: =stdev(B3:B5)

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		15.43333	30.9
Row 7		**= stdev (B3:B5)**	

Parentheses must be used carefully, so that Excel will carry out the operations in the right order, and this is especially important in more complicated formulas. To calculate a raw effect size (specifically, a repeated measures ES using the SD of the pretreatment scores), the following formula is typed into Cell B8:

=(C6-B6)/B7

Row 1	Column A	Column B	Column C
Row 2		Intake Score	Termination Score
Row 3	Client 1	14.2	36.7
Row 4	Client 2	9.7	28.7
Row 5	Client 3	22.4	27.3
Row 6		15.43333	30.9
Row 7		6.43920	
Row 8		**2.40195**	

This result is the difference between the intake and termination scores, divided by the standard deviation of the intake scores, just as required by the raw effect size formula.

The means and SD do not need to be calculated in separate cells (e.g., in Cells B6, C6 and B7 above) in order to calculate an ES. The entire ES formula can be typed within a single cell if the proper syntax is used and parentheses are correctly placed:

=(average(C3:C5)-average(B3:B5))/stdev(B3:B5)

Using the effect size for benchmarking: Effect sizes can be influenced by random variations in a clinician's caseload, so they are likely to be "unstable" or unreliable with caseloads of fewer than 30 clients. Caseloads of 60 or more clients are likely to yield effect sizes that are stable unless systematic changes in therapist functioning or caseload occur; and caseloads of 100 or more clients will provide especially robust or predictive effect sizes (again, unless substantial changes occur in the clinician's overall performance or caseload characteristics). For raw ES calculations, the closer a clinician's caseload mirrors the pretreatment mean and standard deviation (and other client characteristics) of published statistics, the more the clinician can rely on the raw ES as a good estimate of effectiveness, unbiased by an unusual or quirky sample. Typical ranges of pretreatment means for the ORS are approximately 18-19, and standard deviations typically range between 6.5 and 7.5.

ES has often been considered a "unit-less" and standardized measure of effect, to help with comparisons between different outcome measures with different ranges of scores. Yet, different outcome measures can have different sensitivity-to-change or measure different ranges of well-being, leading to different ESs, depending on the measure used. For example, in a study of arthritis treatment on client well-being, effect sizes from different measures of "pain severity" ranged from 0 to 0.9 and measures of "function" or "doing things" ranged from 0 to 0.5. Clinicians are advised against cross-measure

comparisons of well-being ESs until further research provides adequate data to support the comparability of effect sizes yielded by different outcome measures.

Interestingly, studies of clinical change in mental health settings have consistently shown that the greater a client's distress is at the start of treatment, the more he or she is likely to show change by the end of treatment. Conversely, the more well-being a client has at the start of therapy, the less likely he or she will show a large amount of change by the end of therapy. This predictable finding has led some researchers to develop a new kind of effect size that takes into account the severity of a client's distress at the start of therapy. Such an adjustment provides an assessment of the size of the change relative to the client's functioning at the outset. In other words, a client with severe distress at the start of therapy who makes a modest amount of improvement – less than the average expected amount for that degree of distress – would have the change score for that client adjusted lower. A client starting therapy already feeling good who achieved a modest amount of improvement – but more than the expected amount of change for clients with low distress – would have the change score for that client adjusted higher. This kind of effect size in which each client's change score is adjusted based on his or her amount of change compared with the expected amount of change for the initial severity is often called a severity-adjusted effect size (sometimes referred to as "case-mix-adjusted" effect size), and it is used to make fairer judgments about how statistically "easy" or "difficult" it is to show change given the different clinical distress that clients have at the start of therapy.

Severity-adjusted effect size formulas (to adjust an individual therapist's or agency's raw effect size) often are created from very large samples of clients. For an agency that has many clients with above-average or below-average distress at intake, a severity-adjusted effect size provides a more accurate sense of the meaningfulness of clients' change. At the level of an individual clinician, the severity-adjusted effect size can give a fairer estimate of effectiveness relative to one's caseload and other therapists, and can also help clinicians identify whether they are relatively better or worse than expected at helping either clinically distressed or nonclinically distressed clients so that the clinicians can more effectively focus their efforts at improving their service or so they can mentor others to improve their outcomes.

A formula that is similar to the severity-adjusted effect size (using the same kind of linear regression model) can be used to estimate an "expected trajectory of change" (or ETC) for any given intake score. An ETC (discussed in Section 2) is a line graph that shows how much change an average client is likely to experience in each of the first six or seven sessions based on his or her first session score. The ETC is an average based on a large number of clients who have started with a wide range of intake scores, and takes into account that clients with severe distress

at intake show early rapid change, and clients who start therapy with mild distress (on average) show relatively little change. The ETC is a benchmarking tool to help clinicians and administrators know whether a given client is tending to show more or less change than the average client early in treatment.

Another benchmark is the use of an expected termination score or benchmark score based on the client's intake score (a severity adjustment similar to the ones used in the ETC or the severity-adjusted effect size). The expected termination score does not take into account how many sessions the client has received but instead provides a "baseline" or "benchmark" for effectiveness that is the average effectiveness for therapists in a normative sample that was used to create the formula. Each client who finishes therapy can then be described as showing an outcome that was above or below the baseline or benchmark amount of change that would be expected from therapy given his or her severity of distress at intake. The expected termination scores can then be used to calculate the percentage of clients who have terminated therapy above or below this benchmark (mathematically, 50% of all clients would be expected to score higher – and 50% lower – than the benchmark score for any intake score, so a therapist's effectiveness can be judged in comparison to whether more or less than 50% of his or her clients reach the benchmark for their intake score.

Yet another alternative to using the more complicated severity-adjusted effect size is to use the clinical cutoff (see Section 3 of this manual) and measuring the raw effect size only for the clinically distressed clients, and not clients who had a level of well-being at intake that was above the cutoff, and thus similar to the nonclinical general population. The resulting effect size from this distressed sub-sample will be higher than a raw effect size that includes the clients with greater well-being (as many as a third of a typical caseload can contain clients who start therapy in the nonclinical range), and may be more comparable to research studies that only accept clients with significant clinical distress when measuring the impact of a treatment. The purpose of using this kind of effect size is to provide a metric for how effective a clinician is specifically with the clients who are most in need of services.

Table 7: Common measurements of change in psychotherapy

Statistic	Definition	Commonly expressed as:
Reliable Change Index (RCI)	A change in a score that is probably more than a statistical error; calculated at 5 points for the ORS by previous authors	Number of points on a measure; "percent of clients improved or worsened"
Clinically Significant Change (CSC)	Achieving a reliable change that is also a change from a score within the clinical range to a score in the normal range (beyond the clinical cutoff score of 25 on the ORS)	Number or percentage of clients achieving it; "percent of clients recovered or significantly deteriorated"
Effect Size (ES; within-group or repeated-measures)	The difference between scores, based on the variation in the intake scores; also called "the standardized mean difference" between scores	Percentage of a standard deviation, expressed in decimals; e.g., "1.12"
Severity-Adjusted (or Case-Mix Adjusted) Effect Size (SAES)	An effect size that biases each client's amount of "raw" change by the severity of the intake score; expected amounts of change influence the final effect size	Percentage of a standard deviation expressed in decimals; e.g., "0.74"
Expected Trajectory of Change (ETC)	Scores that are predicted to occur at Session 2, 3, etc., based on a given intake score and a linear-regression formula calculated from a large reference group of therapy clients	Line graph used as a benchmark for a client's data; see Figures 1 and 2
Benchmark Score, Baseline Score or Expected Termination Score	The termination score that is most likely to occur for a given intake score based on a large reference group's amount of therapeutic change for varying levels of distress at intake	Number or percentage of clients achieving it; e.g., "59% above benchmark at termination"

Summary

The purpose of collecting outcome data is to help clinicians know whether they are helping their clients so that they can make any necessary adjustments to their current work and also address their overall performance to achieve excellence in their professional life. As they collect data with their clients, clinicians must know how to calculate the basic statistics that will tell them something meaningful about the change or lack of change their clients are experiencing so they can know when it is time to alter their approach. The clinician must:

- establish a valid baseline with standardized measures that have validity and reliability while still being feasible to use at every session, being sure not to miss the essential first-session (baseline) measure of well-being;
- learn how to graph results and appropriately use and explain any signals or benchmarks that are displayed on the graphs; whether with paper and pen, a simple spreadsheet program like Microsoft Excel or sophisticated outcomes management software;
- understand clinical significance, the difference between reliable change and recovery, and some of the assumptions, pros, and cons of using clinical significance statistics; and
- know how to calculate a repeated-measures effect size, understand how expected trajectories of change and severity-adjusted effect sizes are used, and some of the assumptions, pros, and cons of using effect size statistics.

The world of outcome statistics is vast and changing. As with the study of any new material, becoming comfortable and familiar with the basics (as presented here) creates a foundation for further exploration into more advanced areas. It is helpful to reread this material as many times as needed until it becomes clear and familiar. Take the short quiz at the end of this manual, notice what was unclear and then reread the manual after a short break from it. The deliberate, focused engagement with difficult material – again and again until it becomes natural and internalized, and then pressing on to new material to master – is how excellence is built.

Manual 4 Quiz

Research indicates that people retain knowledge better when tested. Take a few moments and answer the following 10 questions. If you miss more than a couple, go back and reread the applicable sections. One week from now, complete the quiz again as a way of reviewing and refreshing what you have learned.

1. **The relationship between reliable change (RC) and clinically significant change (CSC) is that:**
   a. RC is a necessary part of CSC
   b. RC requires measuring reliability, CSC requires asking the client to judge the importance of change
   c. RC involves crossing the clinical cutoff, and CSC can occur without crossing the cutoff
   d. The percentage of clients with CSC is always smaller than the percentage of clients with RC

2. **It is important to assess a measure's validity because:**
   a. Homemade measures cannot accurately measure real and relevant clinical distress
   b. An outcome measure should be sensitive to meaningful change but less sensitive to smaller changes
   c. It will be easier and faster to administer to clients
   d. Validity increases interrater reliability, and high interrater reliability is an element of validity

3. **Cronbach's alpha is a measure of:**
   a. Test-retest reliability
   b. Construct validity
   c. How similarly a client answers the different items on a measure
   d. The deviation of the reliable change from the expected amount of change

4. **The typical relationship between distress at intake and how much change clients experience is:**
   a. The higher the distress at intake, the lower the distress at termination
   b. The higher the distress at intake, the greater amount of change at termination
   c. The lower the distress at intake, the higher the effect size is likely to be
   d. All of the above

5. **Effect size provides the _____; clinically significant change provides the _____:**
   a. Preferred statistic for outcomes; informal statistic for outcomes
   b. Magnitude of change; number of people changed
   c. Standard error; benchmark score
   d. None of the above

6. **Effect size requires a _____; clinically significant change requires a _____:**
   a. Sample size of 200 or more; sample size of 100 or more
   b. Nonclinical sample; inpatient normative sample
   c. Reliable change index; pretreatment mean
   d. Standard deviation and two means; reliable change index and clinical cutoff

7. **The effect size of a clinician's ORS scores is comparable to:**
   a. The ES of other outcome instruments with similar characteristics such as the OQ-45.2
   b. The ES of other clinicians' ORS scores, as long as all clinicians have sample sizes of 30 clients
   c. The ES of other clinicians' ORS scores, as long as all clinicians have similar caseloads and settings
   d. The ES of other clinicians' ORS scores, as long as all clinicians have similar caseloads and settings, large sample sizes, few missing data, careful administration and even then with caution

8. **The expected trajectory of change (ETC):**
   a. Provides a guideline for interventions and conversations about change or lack of change early in treatment
   b. Can provide specific decision-rules for when to terminate, transfer or increase the frequency of sessions
   c. Provides long-term clients with feedback about the likely amount of change late in treatment given a particular level of distress at intake
   d. Can accurately predict the individual client's trajectory of change from session-to-session near the start of treatment

9. **The clinical cutoffs for adults, adolescents and children on the ORS and CORS are:**
   a. 36, 32, 25
   b. 25, 28, 32
   c. 5, 25, 36
   d. 25, 25, 25

10. **According to references cited in this manual, treatment alliance scores on the SRS and other alliance measures near the start of treatment correlate with eventual clinical outcome at termination with a coefficient of approximately**
   a. .25 to .30
   b. .45 to .50
   c. .60 to .65
   d. .75 to .80

Answer Key

1. a	6. d
2. b	7. d
3. c	8. a
4. b	9. b
5. b	10. a

FAQ

Question:

If my outcomes are not that great, should I be worried? What can I do about it?

Answer:

If you are using a valid, reliable outcome measure that is relevant to the clients with whom you work, then positive outcomes should certainly be your goal. Many factors influence the outcomes you obtain, and poor outcomes should lead you to ask: "What is contributing to the lower results that my clients are getting and what can I do about it?" If you are measuring the clinical alliance, the alliance scores may point to some problems that need attention (e.g., in making sure that you and your clients are on the same page or that you quickly and successfully repair ruptures). It is also possible that the fit between you and the setting you work in, or the way clients are screened and matched to you, may need to be changed. It is an excellent idea to seek an independent reanalysis of your data if you have doubts about the accuracy of your outcome statistics, as well as consultation or supervision to understand what may be hindering your clients' experiences of change on the measures you use.

Question:

If my outcomes are especially good, should I be cautious about getting too excited and "believing my own press"? Can I compare my effect size or other outcome statistics with the statistics of other therapists?

Answer:

You should be cautious. There are several reasons to be careful about getting too excited or blowing your own horn when you have excellent outcomes, and there are ethical considerations in the way you report your results. For example, if you have a small sample size (e.g., fewer than 30 clients) it is likely that with more clients your results could change substantially. This concern dwindles if you have a sample of more than 100 clients. If your results involve clients who you worked with in one setting, this does not mean that clients you work with in another setting (who may have different reasons for starting therapy or who may come from a different population or with whom you may interact differently) will achieve the same overall results. On the other hand, research does show that therapists' overall results over an extended period of time tend to be stable year-over-

year (in the same setting). Along the same lines, different clinicians and different agencies may work with different kinds of clients due to differences in intake procedures, screening, mandates, marketing, etc. Statistical differences in the samples can lead to an apples-to-oranges comparison rather than an apples-to-apples comparison; so direct comparisons between clinicians and agencies should be made cautiously and with appropriate caveats about factors that might affect the validity of the comparison.

Question:

Can I compare my effect size (or other outcome statistics) with the effect sizes I read about in research articles? What about comparing with effect sizes that others have reported with measures that are different from the one that I use?

Answer:

Outcomes research often involves substantial differences in design from a simple pre-post measurement of change that most clinicians use in measuring their outcomes. Therefore, the effect size reported in a research article may be based on a different kind of calculation (e.g., a difference between a treatment group and a waiting list group rather than a difference between the first session and the last session for a single group of clients). Researchers may use different methods for recruiting or screening participants in outcome studies than clinicians would use for starting therapy with clients. These differences as well as other potential differences in settings, expectations, etc., may make it difficult to draw simple comparisons between clinical outcomes in "the real world" with the outcomes reported by researchers. One should also be very cautious about making comparisons between the effect sizes or other outcome statistics of two different outcome measures. Different outcome measures may have different "sensitivities" for measuring changes in well-being or other psychological factors, and may actually measure different constructs even while using a similar label such as "depression," "anxiety" or "well-being."

Question:

What does "percent of clients reaching target" or "percent of clients reaching baseline" mean in discussing the outcome statistics for a given clinician?

Answer:

"Percent of clients reaching target" means the number of clients whose scores improve by the end of treatment at least to a score that is expected given their well-being score in the first session (on occasion, other minor factors or "predictors" besides the intake score may also be used to predict the expected posttreatment score), divided by the total number of clients seen. In addition to "target," the word "baseline" or "benchmark" may be used to denote the score that would be expected on average. Typically, the "expected"score is an approximation based on a very large reference pool of clients who start therapy at varying levels of distress and change to varying degrees by the end of treatment. The trend of change shown at different levels of intake distress is then calculated as an equation and this equation is used to estimate an expected amount of change for a given score at the start of treatment. In this kind of equation, half of clients in the very large reference pool of clients are expected to score below the target and half are expected to score above the target based on any intake score.

Question:

Sometimes people refer to an effect size by translating it into the percentage of clients who are "better off from treatment" versus the average untreated client. Or the average client posttreatment may be better off than some percentage of clients who are untreated. What exactly does all this mean?

Answer:

"Untreated" in the case of clinical data comparing first-session and last-session scores refers to clients in their first session. There is a simple way to translate outcomes into a percentage of clients who have greater well-being than the average client did prior to treatment. (1) Calculate the average pretreatment score. (2) Count how many clients' posttreatment scores are above the pretreatment average. (3) Divide that number of clients by the total number of clients. The result is the percent of treated clients who score above the average pretreatment score. There is a mathematical way

of converting an effect size into this kind of percentage but it is advisable to avoid it because the formula for doing so relies on some assumptions that may not be accurate in your sample, and it is also relatively easy to directly calculate the figure without using the effect size.

Question:

Why is the pretreatment standard deviation (SD) used in calculating simple effect sizes?

Answer:

The SD is a measure of the amount of variation or "noise" in a group of scores. One argument for using the SD of the first-session or "pretreatment" scores rather than some other basis for the SD (such as the SD of the change or the SD of the posttreatment scores), is that the effect size should be calculated based on a specific aspect of the variation in scores: the spread in scores prior to any intervention. In other words, the SD used for calculating the effect size should account for how much of the change is due simply to the variation that exists between people at the outset of treatment. Being in treatment typically increases the spread of scores (the variation between scores increases because many clients feel substantially better from treatment and others may feel worse than they did at the start). The more effective treatment is, the higher the SD is likely to be. The higher SD in the denominator of the effect size would reduce the effect size; but the increased variation that results from effective therapy should not count against the effect size. So, the variation in scores prior to the intervention is used as the measure of variation or "noise" among client scores.

References

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(4), 693-704.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist,* 61(4), 271–285.

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(1), 23-29.

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), 1-9

Duncan, B.L., Miller, S.D., Sparks, J.A., Claud, D.A., Reynolds, L.R., Brown, J., & Johnson, L.D. (2003). The session rating scale: Preliminary psychometric properties of a "working alliance" inventory. *Journal of Brief Therapy,* 3(1), 3-11.

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*(2), 71-87.

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(1), 12-19.

Lambert, M.J., Morton, J.J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R.C., Shimokawa, K., Christopherson, C., & Burlingame, G.M. (2004). *Administration and Scoring Manual for the OQ-45.2.* American Professional Credentialing Services, L.L.C. Available at: http://www.oqfamily.com

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(1), 5-22.

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(2), 91-100.

Orlinsky, D.E., Rønnestad, M.H., Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.)(pp. 307-390). New York: Wiley.

INTERNATIONAL CENTER FOR CLINICAL EXCELLENCE