The contribution of reliable and clinically significant change methods to evidence-based mental health

Where outcomes are unequivocal (life or death; being able to walk *v* being paralysed) clinicians, researchers, and patients find it easy to speak the same language in evaluating results. However, in much of mental health work initial states and outcomes of treatments are measured on continuous scales and the distribution of the "normal" often overlaps with the range of the "abnormal." In this situation, clinicians and researchers often talk different languages about change data, and both are probably poor at conveying their thoughts to patients.

Researchers traditionally compare means between groups. Their statistical methods, using distributions of the scores before and after treatment to suggest whether change is a sampling artefact or a chance finding, have been known for many years.[1] By contrast, clinicians are more often concerned with changes in particular individuals they are treating and often dichotomise outcome as "success" or "failure." The number needed to treat (NNT) method of presenting results has gone some way to bridge this gap but often uses arbitrary criteria on which to dichotomise change into "success" and "failure." A typical example is the criterion of a 50% drop on the Hamilton Depression Rating Scale score. A method bridging these approaches would assist the translation of research results into clinical practice.

Jacobson *et al* proposed a method of determining **reliable and clinically significant change (RCSC)** that summarises changes at the level of the individual in the context of observed changes for the whole sample.[2 3–5] Their methods are applicable, in one form or another, to the measurement of change on any continuous scale for *any* clinical problem, although they have been reported primarily in the psychotherapy research literature.

The broad concept of reliable and clinically significant change rests on 2 questions being addressed at the level of each individual subject:

- *Has the patient changed sufficiently to be confident that the change is beyond that which could be attributed to measurement error?*—**"reliable change"** *and*
- *How does the end state of the patient compare with the scores observed in socially and clinically meaningful comparison groups?*—**"clinically significant change."**

We address each of these components, a range of issues relating to these procedures, and conclude with a number of recommendations for practitioners and researchers.

Reliable change: what is it and how is it calculated?

Reliable change refers to the extent to which the change shown by an individual falls beyond the range which could be attributed to the measurement variability of the instrument itself. The measurement variability is termed the Reliable Change (RC) Index and is assessed using a variation on the standard error (SE) of measurement which takes account of 2 measurements being made (before and after treatment). This is called the SE of the difference.[2–5] The formula for the SE of measurement of a difference is:

$$SE_{diff} = SD_1\sqrt{2}\,\sqrt{1-r}$$

where SD_1 is the standard deviation of the baseline observations and, r is the reliability of the measure.

Change exceeding 1.96 times this SE is unlikely to occur more than 5% of the time by unreliability of the measure alone. If the internal reliability, Cronbach's coefficient alpha (α), is used then the methods arise straightforwardly out of classical measurement theory. Generally, test-retest reliabilities are lower than α because they contain an additional source of variation: any real changes in the variable of interest. Use of any other test-retest reliability creates confusion about what is error of measurement and what is real change.

Clinically significant change: what is it and how is it calculated?

A separate consideration from the reliability of the change is the extent to which change after treatment is clinically meaningful. The original principle was to judge the change against socially validated criteria: Jacobson *et al* suggested that clinical significance (what *Evidence-Based Mental Health* refers to as "clinical importance") was represented by a person's score moving from the "dysfunctional population" range into the "functional population" range.[2] They suggested 3 criteria (termed A, B, and C) to be based on the scores of the change variable:

(A) Pre-change to post-change of at least 2 standard deviations (SDs) from the original mean. This does not fulfil the aim of comparing the end point to a reference "normal" population.

(B) Change moving the patient to within 2 SDs of a normative sample mean. This does not measure the extent to which the patient is also moving out of the original sample.

(C) Greater likelihood of the patient being in the normative distribution than a clinical distribution after treatment. This requires determination of the "cut off" point where the probability of coming from each of the distributions is equal. Where the SDs of the clinical and normative data are equal, this is the mean of the first 2 criteria; more generally it is:

$$\frac{(mean_{clin} \times SD_{norm}) + (mean_{norm} \times SD_{clin})}{SD_{norm} + SD_{clin}}$$

Figures 1 and 2 show these 3 criteria. As can be seen in figure 1, when distributions are overlapping, criterion A can seem stringent and criterion B lenient. In that situation, criterion C is generally best. When the distributions are essentially non-overlapping (figure 2), B is stringent but the most credible criterion if the aim is that treatment should return individuals to the "normal" or "general" population range. Here, even C may still be too stringent a criterion to pick up meaningful, clinically worthwhile change falling short of

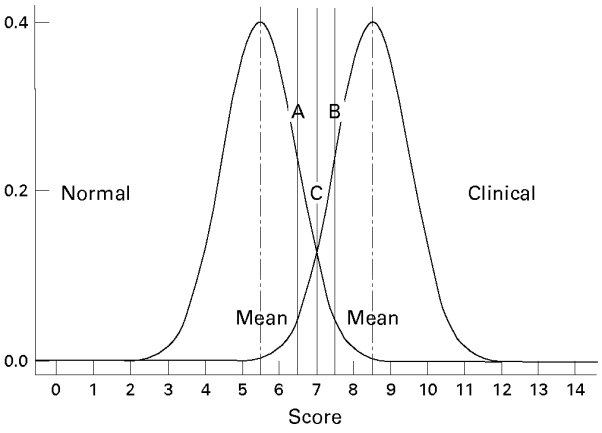

Figure 1 Distributions with criteria for clinically significant change: overlapping.

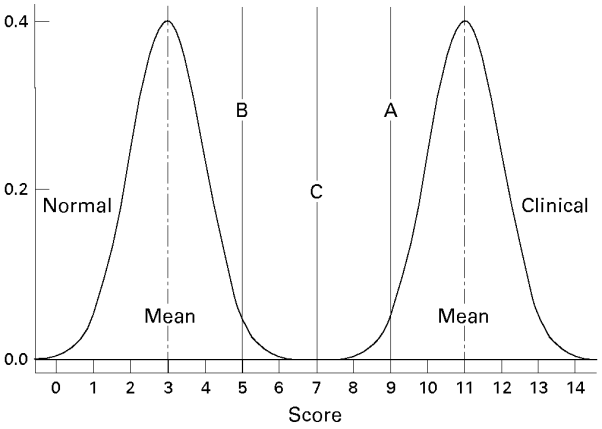

Figure 2 Distributions with criteria for clinically significant change: non-overlapping.

"cure." In this situation the reliable change criterion gains greater importance.

For severe problems normative data from "well" or "general population" samples may be replaced with less disabled clinical samples, for example outpatient or primary care samples. These could form a "cascade" of different distributions which refer to socially validated population samples chosen appropriately in the light of the intentions of the clinicians and service.[6]

A worked example

As part of a continuing project, self report data were collected before and after very short term psychological intervention in a university student counselling service. The measure was the CORE 34 item measure covering symptoms, functioning, wellbeing, and risk to self and others.[7] The 40 students who had complete before treatment and after treatment data showed a before treatment mean (SD) score of 1.81 (0.53) and an after treatment mean (SD) of 0.79 (0.50). The distributions did not differ significantly from a Normal one and using conventional group mean methods the paired t test showed a significant change ($t(39) = 9.5$, $p < .0005$). Coefficient α for the 34 items was 0.89, a little lower than in other large samples for this measure but very respectable.[7]

The SE of measurement of the difference is:

$$SE_{diff} = .53\sqrt{2}\,\sqrt{1-.89} = .53*1.414*.332 = .249$$

hence change that exceeds $1.96 \times 0.249 = 0.487$ can be regarded as reliable. Inspecting the data, 9 of the 40 students

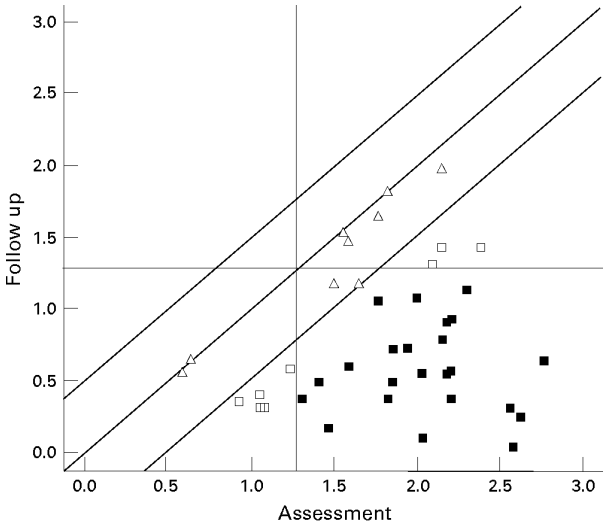

Figure 3 Scatterplot of assessment/follow up scores showing reliable and clinically significant change criteria.

(22.5%) showed change smaller than this. The lowest change was zero (for 2 students) so there were none with reliable deterioration and 77.5% showed reliable improvement.

A referential student volunteer sample (n = 686) from a not dissimilar university showed a mean (SD) score of 0.72 (0.57). This gives the various clinically significant change criterion scores as follows:

Criterion A (initial mean minus twice the SD):
$1.81 - (2) * 0.53 = 0.75$.

Criterion B (normative mean plus twice that SD):
$0.72 + (2) * 0.57 = 1.86$.

Criterion C:

$$\frac{mean_{clin}SD_{norm} + mean_{norm}SD_{clin}}{SD_{norm} + SD_{clin}} = \frac{1.81*.57 + .72*.53}{.57 + .53} = \frac{1.03 + .38}{1.1} = 1.28$$

C is the most appropriate criterion because we have referential data and moving to a greater probability of being in the "normal" population is an appropriate aim. By this criterion, 7 (17.5%) started below the criterion so could not show clinically significant change, and 5 (12.5%) of these 7 showed reliable improvement. Eight (20%) started above the criterion but failed to improve to below it, of these 3 (7.5%) showed reliable improvement. Finally, 25 (62.5%) showed clinically significant change, of whom 23 (57.5%) also showed reliable improvement. The table summarises these data.

The calculations are spelt out at length here but are trivial with a calculator. We do not recommend applying the first 2 criteria to such data but the values are calculated to help those who want to be sure they follow the method. For those who prefer to have things done for them, forms to do all the calculations are available.[8] An SPSS program and SAS program which uses the criteria to produce frequency tables and cross tabulations such as that in the table can also be

Cross tabulation of reliable change against clinically significant change (criterion C)

Clinically significant change (criterion C)	*Reliable change*		
	Yes	*No*	*Total*
Failed to achieve clinically significant change despite sufficient initial score	5	3	8
Started better than criterion for clinically significant change	2	5	7
Clinically significant change	2	23	25
Total	9	31	40

found there, as well as an S-plus program to plot the results as figure 3 shows.

Figure 3 presents a visual display of before intervention and after intervention data. Each point is a student: the x axis is the assessment (pre-treatment) score on the measure, and the y axis is the follow up score. Thus, points lying on the diagonal showed no change, those above it showed deterioration, and those below showed improvement. The horizontal and vertical marker lines show criterion C in relation to assessment and follow up. The solid squares denote data points (individuals) achieving reliable change (ie, change beyond the bounds of measurement error determined using coefficient α) and clinically significant change (ie, post-treatment score below the cut off point based on criterion C). By contrast, the unfilled squares denote individuals whose change data are reliable but do not show clinically significant change. Any data points falling within the reliability "tramlines," as denoted by the unfilled triangles, show change that could be attributable to error measurement.

What problems are there with these methods?

It has been shown that the use of a cut off point will not be estimated well by the conventional cut off point C, described originally by Jacobson *et al* when the distributions have different variances and skew.[9] For severely skewed distributions that cannot be normalised, these methods will not have the simple interpretations in terms of percentages given above.

Regression to the mean is always important. To the extent that people enter treatment at distressed times, rather than non-distressed times in their lives, there will be a systematic regression to the mean whether or not they receive treatment. Services offering crisis intervention are likely to show more such drop in scores, services with long waiting times will show less. However, this cannot be separated from treatment effects unless multiple observations are made before clinical intervention. Regression to the mean affects reliable change and clinically significant change. Suggestions have been made to correct for this,[10 11] but given uncertainties about normal fluctuations over time we believe these replace a simple, if biased, value with one that is almost certainly still biased and not at all simple. For research purposes where multiple observations can be made over time, there are a number of statistical procedures (eg, multilevel or random coefficients regression models) that will provide more detailed and robust methods than RCSC and are preferable to modifications to the RCSC formulas. However, for the clinician or department wishing to provide a personal or internal evidence base for evaluation of effectiveness, RCSC methods will be unlikely to be displaced by those highly complex and computationally demanding methods.

Recommendations

- Reliable and clinically significant change should be reported in articles to complement the more familiar group summary methods.
- Where NNTs are reported they should be based on reliable or clinically significant change (or both).
- RCSC should be used in service evaluation to allow comparisons between services and even between practitioners *at the level of individual patient change*. The methods systematically identify good and bad outcomes (outliers) for individual case audit.
- Comparisons will be easier if a few reference instruments with data on a range of normative samples are used. We have argued elsewhere for a routine "core battery" which can be used in service settings as well as formal outcome research.[12]
- Where non-clinical referential data are used to determine the reliability coefficient and "cut off" points, these should be relevant to the study sample. The referential data should be gathered in the same country (or at least a similar one), should be for similar age distributions, and take into account comparability of sex, socioeconomic status, and ethnicity. The source and actual values should be clearly stated.
- Where clinical referential data are used, the level of severity and clinical location of the samples on which they were derived should be described briefly as well as the demographic variables mentioned earlier. Again, actual values should be clearly stated.
- The basis for the calculation of reliable change should be stated with justification for the choice between α coefficient and test-retest reliability. The source and value of the reliability figure used should be given.

CHRIS EVANS
St Georges Hospital Medical School, Cranmer Terrace, London, SW17 ORE, UK
FRANK MARGISON
University of Manchester, Manchester, UK
MICHAEL BARKHAM
University of Leeds, Leeds, UK

1 Hill AB. *Principles of medical statistics.* London: Lancet 1937.
2 Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy* 1984;**15**:336–52.
3 Christensen L, Mendoza JL. A method of assessing change in a single subject: an alteration of the RC index. *Behavior Therapy* 1986;**17**:305–8.
4 Jacobson NS, Revenstorf D. Statistics for assessing the clinical significance of psychotherapy techniques: issues, problems, and new developments. *Behavioural Assessment* 1988;**10**:133–45.
5 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;**59**:12–9.
6 Tingey RC, Lambert MJ, Burlingame GM, *et al.* Assessing clinical significance: proposed extensions to method. *Psychotherapy Research* 1996; **6:**109–23.
7 CORE System Group. *CORE system (information management) handbook.* Leeds: CORE System Group, 1998.
8 Evans C. *Reliable and clinically significant change.* http://psyctc.sghms.ac.uk/stats/rcsc.htm, 1998.
9 Martinovich Z, Saunders S, Howard K. Some comments on "assesing clinical change." *Psychotherapy Research* 1996;**6**:124–32.
10 Hageman WJ, Arrindell WA. A further refinement of the reliable change (RC) index by improving the pre-post difference score: introducing RCID. *Behav Res Ther* 1993;**31**:693–700.
11 Hsu LM. Reliable changes in psychotherapy: taking into account regression toward the mean. *Behavioral Assessment* 1989;**11**:459–67.
12 Barkham M, Evans C, Margison F, *et al.* The rationale for developing and implementing core outcome batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health* 1998;**7**:35–47.

OTHER ARTICLES NOTED

The journals that are reviewed and the criteria for selecting articles from these journals for inclusion in *Evidence-Based Mental Health* are set out in the purpose and procedure in each issue. All articles that meet our criteria in the reviewed journals are cited in *Evidence-Based Mental Health*, but there is not enough space to abstract them all. The following articles passed all criteria but were not abstracted because, in the judgment of the editors, their findings were less widely applicable to clinical practice in the area of mental health.

Therapeutics

Metrifonate treatment of the cognitive deficits of Alzheimer's disease. Cummings JL, Cyrus PA, Bieber F, *et al. Neurology* 1998 May;**50**:1214–21.

CDP-choline in the treatment of cognitive and behavioural disturbances associated with chronic cerebral disorders of the aged. (Cochrane Review, latest version 09 January 1998). Fioravanti M, Yanagi M. In: the Cochrane Library. Oxford: Update Software.

Randomised controlled trial of compliance therapy: 18-month follow-up. Kemp R, Kirov G, Everitt B, *et al. Br J Psychiatry* 1998 May;**172**:413–9.

Donepezil improves cognition and global function in Alzheimer disease: a 15-week, double-blind, placebo-controlled study. Rogers SL, Doody RS, Mohs RC, *et al*, and the Donepezil Study Group. *Arch Intern Med* 1998 May 11;**158**:1021–31.

Mirtazapine: efficacy and tolerability in comparison with fluoxetine in patients with moderate to severe major depressive disorder. Wheatley DP, van Moffaert M, Timmerman L, *et al*, and the Mirtazapine-Fluoxetine Study Group. *J Clin Psychiatry* 1998 June;**59**:306–12.

Diagnosis

Detecting psychiatric morbidity after stroke: comparison of the GHQ and the HAD Scale. O'Rourke S, MacHale S, Signorini D, *et al. Stroke* 1998 May;**29**:980–5.

Aetiology

Rapid tryptophan depletion, sleep electroencencephalogram, and mood in men with remitted depression on serotonin reuptake inhibitors. Moore P, Gillin C, Bhatti T, *et al. Arch Gen Psychiatry* 1998 June;**55**:534–9.

Effects of tryptophan depletion vs catecholamine depletion in patients with seasonal affective disorder in remission with light therapy. Neumeister A, Turner EH, Matthews JR, *et al. Arch Gen Psychiatry* 1998 June;**55**:524–30.

Prognosis

Psychosocial predictors of functional change in recently diagnosed rheumatoid arthritis patients. Evers AWM, Kraaimaat FW, Greenen R, *et al. Behav Res Ther* 1998 Feb;**36**:179–93.

A prospective study of heart rate response following trauma and the subsequent development of posttraumatic stress disorder. Shalev AY, Sahar T, Freedman S, *et al. Arch Gen Psychiatry* 1998 June;**55**:553–9.

Erratum: in the August 1998 issue of *Evidence-Based Mental Health* there is an editorial error in the table in the second notebook article. Column two should be headed "No" and column three should be headed "Yes" (as shown below).

Cross tabulation of reliable change against clinically significant change (criterion C)

	Reliable change		
Clinically significant change (criterion C)	*No*	*Yes*	*Total*
Failed to achieve clinically significant change despite sufficient initial score	5	3	8
Started better than criterion for clinically significant change	2	5	7
Clinically significant change	2	23	25
Total	9	31	40