INVITED PAPER

Reliable change and the reliable change index: still useful after all these years?

Neville M. Blampied

University of Canterbury, Christchurch, New Zealand
Email: Neville.blampied@canterbury.ac.nz

Abstract

In 1984 Jacobson and colleagues introduced the concept of reliable change, viz the amount of change on a measure that an individual needed to show to determine that it exceeded the extent of change likely due to measurement error alone. Establishing reliable change was a pre-requisite for determining clinical significance. This paper summarizes the rationale for determining reliable change as providing an individual-focused, idiographic alternative to the dominant nomothetic approach to clinical outcome research based on group mean data and statistical significance. The conventional computational steps for calculating an individual's standardized difference (reliable change) score and the minimum raw change score on the measure (a reliable change index) required to classify individuals as reliably positively changed, indeterminate, or reliably deteriorated are described. Two methods for graphically representing reliable change are presented, and a range of possible uses in both research and practice settings are summarized. A number of issues and debates concerning the calculation of reliable change are reviewed. It is concluded that the concept of reliable change remains useful for both cognitive behavioural researchers and practitioners, but that there are options regarding methods of computation. In any use of reliable change, the rationale for selecting among method options and the exact computations used need clear and careful description so that we can continuously judge the utility and appropriateness of the use of reliable change and enhance its value to the field.

Key learning aims

(1) Recognizing why the concept of reliable change and the reliable change index is still important.
(2) Understanding the conventional formulas for calculating reliable change and the reliable change index (the Jacobson-Truax (JT) method).
(3) Seeing key ways that both researchers and practitioners can use reliable change to improve both research and practice.
(4) Understanding how several issues and debates that have arisen concerning the estimation of reliable change (e.g. how to accommodate practice effects) have progressed.
(5) Recognizing that there are a range of ways that reliable change may be estimated, and the need to provide full details of the method used in any particular instance of its use.

Keywords: clinical significance; Jacobson-Truax method; measurement error; reliable change; reliable change index

Introduction

The study of change is profoundly important to psychology in general, to all clinical interventions, and thus, to cognitive behavioural research and practice. The changes studied may, for instance, be

of changes in behaviour, emotion and cognition as they develop over the lifespan; change may be due to the operation of psychopathological processes or be the positive (or negative) outcome of therapy. Change may occur for good or ill as a consequence of social influences; it may signal the enhancement of or deterioration in the performance of individuals, groups and organizations; and involve changes in performance due to learning and instruction. In all research, observed changes capture the myriad outcomes of observation and experimentation across all domains of psychological investigation. In all cases the detection of change requires repeated measurement of one or more dependent variables (a minimum of two measurement points per variable) from a particular case. The cases studied are typically an individual, but may be larger entities, such as a family, a social group, a school class, a sports team, a work group, a company or an organization, or other coherent bodies (Valsiner, 1986).

For the past 70+ years, the study of change in psychology has been dominated by an approach to science that the German philosopher Windelband termed 'nomothetic' (Lamiell, 1998; Windelband, 1894/1998); such an approach is concerned with establishing the general laws of the science. The way that psychology has investigated change within this framework has involved the (quasi-random) allocation of participants to groups, with some treatment hypothesized to induce change applied to one or more groups while one or more separate groups do not experience the treatment and are control groups. Individuals' data are aggregated into measures of central tendency such as means, and the question becomes 'Are mean differences between treatment and control groups sufficient to provide evidence of a treatment effect?'. Such a comparison does not, of course, provide direct evidence of change, only of difference. For that there needs to be measurement of the participants before (time 1, t_1) and after (time 2, t_2) treatment in the treatment group and a parallel set of measurements at t_1 and t_2 for the control group. The question then becomes 'Is the evidence of change from time 1 to time 2 in the mean scores of the treatment group sufficiently different from changes in mean scores in the control group across time to provide evidence of a treatment effect?'.

Of course, the interpretation of mean scores across time and between groups is complicated by variance at each measurement point. To detect the signal of a treatment effect from the noise of variance, researchers use statistical techniques developed by R.A. Fisher, termed the analysis of variance (ANOVA; Fisher, 1935) and subsequently extended by many other statisticians (Wright, 2009). Such techniques yield a test statistic (F in the case of ANOVA) and a probability value for that statistic (p) under a null hypothesis (Ho) which is that there is no treatment effect and any observed mean differences are the result of sampling error (Ho_{SampE}). If p is small (conventionally $<.05$, a criterion value referred to as *alpha*; Hubbard, 2004) then Ho_{SampE} can be rejected and the results taken as evidence of a treatment effect. Procedures combining group aggregate mean data and statistical tests that assume Ho_{SampE} are referred to as null hypothesis statistical tests (NHST), and results obtained when Ho is rejected are said to be *statistically significant*, often shortened just to *significant*. Because this categorization of results is based on a probability, it may be wrong. A Type 1 error is the rejection of Ho when it is true, or Ho may not be rejected when it is false, a Type 2 error. Statistical significance has no necessary implications for the clinical or practical significance of the findings and applies only at the aggregate not the individual level.

Psychologists began to use NHST in research in the mid-1930s and since the mid-1950s it has completely dominated research practice (Hubbard *et al.,* 1987). Clinical outcome research quickly followed, as documented by Bergin and Strupp (1970) and Kiesler (1971), and clinical researchers, including those investigating cognitive behaviour therapy (CBT), now almost universally regard the randomized control trial (RCT), invariably analysed by NHST, as the 'gold standard' research practice, clearly endorsed today in the chapters on clinical research methods in the authoritative handbooks edited by Barlow (2014), Norcross *et al.* (2016) and Wright and Hallquist (2020).

Despite this sustained commitment by psychology research, including clinical research, to NHST, criticisms have been persistently voiced (see Balluerka *et al.*, 2005; Nickerson, 2000),

many of these criticisms focusing on the misinterpretation of NHST, especially of *p* values and statistical significance, and the common violation of its assumptions (Cohen, 1990; Cohen, 1994; Colquhoun, 2014; Gigerenzer, 2018; Lambdin, 2012; Schneider, 2015). So persistent did this criticism become that both the American Psychological Association (APA) and the American Statistical Association (ASA) established expert groups to evaluate NHST. Their respective reports (ASA: Wasserstein and Lazar, 2016; APA: Wilkinson and Task Force on Statistical Inference, 1999), while not suggesting the abandonment of NHST, made many recommendations for the improvement of research practices involving it, although there is not much evidence that these have resulted in extensive improvements in practice (Vacha-Haase *et al.*, 2000).

There is an equivalent history of criticism of NHST within clinical research, articulated early by Bergin and Strupp (1972) when they noted that … *statistical procedures … are held inappropriate … to research in the area of therapeutic change* (p. 440). Dar *et al.* (1994) reviewed 30 years of clinical research published in the *Journal of Consulting and Clinical Psychology*, research dominated by NHST, and documented persistent and extensive misinterpretation of the statistics. Fidler *et al.* (2005) duplicated this survey for research in the same journal from 1993 to 2001 and found that NHST-based research continued to dominate, and that misinterpretations and misapplications of statistical methods continued to be common. As, over time, it is the outcome of CBT that has become the predominant research question in clinical research, this history of misuse and misinterpretation of NHST is clearly of concern to researchers and practitioners of CBT.

A central reason why NHST-based research procedures are inappropriate to research into therapeutic change has to do with the central practice of between-subject averaging. This creates a 'double standard' (Valsiner, 1986) where we profess to be concerned with understanding and helping the individual client, but in research, conceal individuals in group averages. Even when a statistically significant therapeutic change is detected at the level of the group means, it is impossible to know what the fate of any individual in the treatment group was, or what the response to treatment might be of any client subsequently given the treatment (Barker *et al.*, 2016; Barlow *et al.*, 1984; Kravitz *et al.*, 2004; Kent and Hayward, 2007). This concern is not confined just to therapy-outcome research, but is a general issue across all research domains (Fisher *et al.*, 2018; Speelman and McGann, 2016).

This concern regarding group averaging has led to appeals that clinical research should become more idiographic (Barlow and Nock, 2009). The term idiographic was also coined by Windelband (Lamiell, 1998; Windelband, 1894/1998) and since being introduced to psychology (along with nomothetic) by Gordon Allport in 1937, idiographic has come to mean the study of the individual over time (Danziger, 1990; Lamiell, 1998), as distinct from the nomothetic investigation of general laws at the group/population level. As Barlow and Nock (2009) noted, clinical science has profited extensively from the nomothetic research tradition, which has laid the foundation for evidence-based practice (Chambless and Hollon, 1998, 2012; Spring, 2007), and so they did not advocate the rejection of such research. Rather, they suggested that we might *enrich these methodologies with a complementary focus on the individual* (Barlow and Nock, 2009; p. 20).

Idiographic methods that might complement traditional NHST research include single-case research designs (Barlow *et al.*, 2009; Blampied, 2013; Cooper *et al.*, 2020) and mixed methods incorporating qualitative and quantitative research methods (Doyle *et al.*, 2009; Levitt *et al.*, 2018; Morgan, 1998); however, these methods will not be considered further here. Rather, I will focus on a specific suggestion, made by Jacobson *et al.* (1984; Jacobson and Truax, 1991) for complementing the NHST-based findings of group outcome research with investigation of individuals' reliable change (RC). The primary purpose of their writing was to introduce standard criteria for the determination of the clinical significance of change by any individual. However, having done so, they noted: *It is nonsensical to speak of clinically*

significant treatment effects when no change has occurred (Jacobson *et al.*, 1984; p. 343). Establishing RC is thus a pre-requisite for determining clinical significance; the measured change must reliably exceed that which might be due just to measurement error alone. Clinical significance remains an extremely important topic (Follette and Callaghan, 2001; Ogles *et al.*, 2021), but it is RC I will focus on in this paper.

Reliable change – conceptual foundations and computational procedures

The presentation below is based on the work of Jacobson and colleagues, especially Jacobson and Truax (1991) that incorporated a formula correction supplied by Christensen and Mendoza (1986). Their work is anchored in classical test theory (Maassen, 2000) and contains an explicit, but different, null hypothesis – the Ho that individual true score differences between t_1 and t_2 are zero and any observed difference is due only to measurement error (Ho$_{ME}$). Jacobson and colleagues also made some further assumptions that underpinned their work. These include that the measure used to assess therapy effects was a valid measure of the attribute of interest, that the attribute of interest was actually targeted by the therapy being investigated, and that the measure was responsive, i.e. that it was sensitive to change. It is also assumed that the conditions under which the measurement is made are optimum for the purpose and that those who administer and score the test are highly competent to do so. These assumptions remain relevant to all subsequent applications of the idea of reliable change.

Reliable change – foundations in classical measurement theory

Classical measurement theory holds that any measurement can be separated into two components – the true score, plus an error component, so that:

$$X = T \pm E \tag{1}$$

where X = the observed score, T = the true score, and E = error (Stigler, 1986). If change has occurred in an individual case over the measured time period in the domain captured by the dependent variable and within the limits of the sensitivity of the particular instrument, then the true score at t_1 will not be equal to the true score at t_2. However, given the variability and error intrinsic to psychological measurement it is possible that a difference between t_1 and t_2 scores may be observed in the absence of any actual change. How then can we determine how large an observed change needs to be to warrant a conclusion that change has truly occurred or decide that our observations are capturing only trivial variation? Enter the concept of RC and the reliable change index (RCI; Jacobson *et al.*, 1984; Jacobson and Truax, 1991).

Deriving RC/RCI depends on also knowing that the frequency distribution of error scores (i.e. where the magnitude of errors of over- and under-estimation is represented on the *x*-axis about a central value of zero and the frequency with which each error has occurred is plotted on the *y*-axis) is represented by the Gaussian distribution, once also known as the Normal Law of Error, and now known (thanks to Galton and Pearson) as the Normal distribution (Stigler, 1986). Importantly, errors of measurement are normally distributed even when the measures themselves are skewed and all the mathematical properties of the normal distribution apply to the error distribution (Stigler, 1986). The mean of the distribution (its peak) is zero, and it has a standard deviation (*SD*), which in the case of the error distribution of measurement errors is given a special name, the standard error of measurement (*SE*$_M$), computed as:

$$SE_M = s_1 \sqrt{1 - r_{xx}}, \tag{2}$$

where s_1 = the standard deviation of a control or pre-treatment group and r_{xx} = the test–retest reliability of the measure.

The SE_M is a measure of the precision of the measurement instrument and formula (2) tells us something that is intuitively reasonable, namely that variation in error of measurement is a function of two things: first, the intrinsic variability of the thing we are trying to measure, and second, the reliability of our measuring instrument. SE_M increases as variability increases and reliability decreases, emphasizing the desirability of using highly reliable measures. From what we know about the normal distribution we also know that, for the error distribution, specified proportions of all the error frequencies will lie within specified values of the SE_M about the mean (zero), such that, for example, 95% of all errors will lie within the range of $\pm 1.96\ SE_M$ and only 5% lie beyond that range.

The standardized change score

Jacobson and colleagues (Jacobson *et al.*, 1984; Jacobson and Truax, 1991) drew on this basic knowledge in order to define RC, but to understand their definition first we have to understand the error distribution of difference scores. Recall, from above, that the basic way of identifying change (if it has occurred) is by a non-zero difference score, in this case a raw change (hence 'C') score, given by the subtraction of the *i*th case's score at t_2 from t_1:

$$C_{i(raw)} = X_{t1} - X_{t2}, \tag{3}$$

where $C_{i\ (raw)}$ is the raw change score for individual i, and X is their score measured at two time points, t_1 and t_2. As each individual measurement contains both a true score and an error score, any raw change score will comprise the true change score $\pm$ an error score (as per equation (1) above), and as the error score is a compound of the errors made at each measurement it will be larger than for the component individual scores. The frequency distribution of difference score errors is also a normal distribution with mean = zero but with an adjusted (larger) standard error, called S_{Diff}, where:

$$S_{Diff} = \sqrt{2(SE_M)^2}. \tag{4}$$

Because all measurement contains error we cannot know what the true absolute scores are or the true change score actually is but we can use what we know about the distribution of errors, and specifically, S_{Diff}, to decide what a reliable change is for any particular measurement, so long as we know SE_M and S_{Diff}. The first step for Jacobson and Truax (1991) in defining RC was to standardize C_i, by dividing the difference score by S_{Diff} :

$$C_{i(Standardized)} = C_i / S_{Diff}. \tag{5}$$

This converts the raw change score to standard deviation (specifically S_{Diff}) units, just as a *z*-score is the difference between an individual's score and the mean standardized by the *SD*.

Interpreting the standardized change score to yield reliable change

The definition of RC using $C_{i\ (Standardized)}$ draws on exactly the same logic as that used in NHST, except that Ho_{ME} is used rather than Ho_{SampE}. We assume Ho_{ME} to be true, and, due to measurement error alone, 95% of all standardized scores will lie within $\pm 1S_{Diff}$. So if $C_{i\ (Standardized)} > 1.96$ or < -1.96 (i.e. it lies at least 1.96 S_{Diff} units either side of the mean) the probability of an error of measurement this large is $p \leq .05$, because only 5% of the frequency distribution of errors lies at this extreme; an error of measurement alone sufficient to produce such a value of $C_{i\ (Standardized)}$ is regarded as unlikely (although not impossible) and Ho_{ME} (the proposition that there is no true change) is rejected (but note that Type 1 and Type 2 errors can still occur). This means that the difference observed probably contains both true change plus measurement error; this is **reliable change** (i.e. change that we can defensibly claim to be larger than that due to measurement error alone). As Jacobson and Truax (1991)

put it: *RC tells us whether change reflects more than the fluctuations of an imprecise measuring instrument* (p. 14). Other standard deviation units can, of course, be used as the criterion to reject Ho, so that $C_{i \text{ (Standardized)}} > 1.645$ or < -1.645 gives $p < .1$, and > 2.58 or < -2.58 gives $p < .01$, etc., but $RC_{(.05)}$ is the conventional criterion (Jacobson and Truax, 1991; note the use of the subscript to specify the probability value, sometimes also given as $RC_{(95\%)}$). Because ±1.96 is regarded as a very stringent criterion for RC some investigators, including neuropsychologists, often use ±1.645 (Duff, 2012).

Thus, for every individual case that we have measured across two or more times in any study we can determine for any pair of measurements if the raw change displayed is large enough to be considered reliable (given that we have the SE_M of our measure). Recall, from above, however, that $C_{i \text{ (Standardized)}}$ may be positive or negative, depending on the values of the raw scores at t_1 and t_2; specifically, if $X_1 < X_2$, $C_{i \text{ (Standardized)}}$ will be negative. In computing RC it is important to retain this information, as it is critical to interpreting RC.

In almost all instances where we measure change we can determine *a priori* the direction of change that is expected or predicted as consistent with theory, normal, beneficial, desirable, or otherwise. For instance, if we are measuring physical growth over time we expect raw scores to increase; if we are measuring psychopathology before and after psychotherapy and our measure of psychopathology gives lower scores when levels of psychological distress are reduced we expect effective psychotherapy to be accompanied by reduced raw scores. So, the directionality of change has meaning, and RC needs to be interpreted in that light. At minimum we can classify RC into three categories: RC that is beneficial is termed Improved (RC+); RC that is non-beneficial is termed Deteriorated (RC–); and where change is insufficient to be classified as reliable, that case is termed Indeterminate (RC0; Jacobson *et al.*, 1984; Wise, 2004). Note that here the sign on the RC indicates the positive/beneficial or negative/deteriorated direction of change, not the arithmetic sign of the change score, so that an individual may show RC+ on two measures of outcome even though one measured improvement in increased scores and the other via decreasing scores. Finer graduations in outcome classification are possible such as the categories of *Improved*, *Remitted*, *Recovered*, each requiring a more stringent RC criterion to be met, as shown in Wise (2004; table 1, p. 56).

The reliable change index – an alternative classification scheme

Following Jacobson and colleagues, RC has been defined above in terms of a standardized change score that is compared against a selected criterion (e.g. ±1.96) derived from the normal curve to give a chosen probability value ($p < .05$ in the case of ±1.96). Because the difference score is standardized it is expressed in deviation units not in the units of the original measure, and is, therefore, measure-independent, like the effect size (ES) Cohen's *d*/Hedges' *g* (which is the group mean difference standardized by the pooled *SD* or the pre–post mean difference standardized by the average *SD*; Borenstein, 2012; Lakens, 2013), and can be used to determine RC across measures with different absolute ranges of values. As an alternative approach to determining RC it is easy to calculate an index score for the particular measure which indicates the absolute value of the difference score required for the difference to be reliable at some criterion level (e.g. $p < .05$).

This index is the RCI and is calculated (specifically for $p < .05$) as:

$$RCI_{(.05)} = 1.96 S_{Diff}. \tag{6}$$

This is generally rounded to the nearest whole number to reflect the units of the measure. Rounding up or down will slightly alter the area under the normal curve that gives the probability value, but given the effects on measurement resulting from forcing individuals to respond using the fixed values of the Likert scales commonly used in psychology research there is no need to be too pious about this. In practice, investigators can use either

Table 1. Raw pre–post scores from 25 cases showing the raw change score, the raw difference score, the standardized difference (change) score, the classification of the change (compared against ±1.96) as reliable improvement (RC+), indeterminate change (RC0), and reliable deterioration (RC–), and the parallel classification using a reliable change index = 5

Case number	Pre	Post	Pre–Post difference	Standardized difference	RC+/RC0/RC–	=/>5 or =/>–5	
1	37	4	33	13.52	✓	Yes	
2	41	29	12	4.92	✓	Yes	
3	20	9	11	4.51	✓	Yes	Reliably
4	23	14	9	3.69	✓	Yes	improved
5	21	12	9	3.69	✓	Yes	
6	9	2	7	2.87	✓	Yes	
7	9	3	6	2.46	✓	Yes	
8	6	1	5	2.05	✓	Yes	
9	19	14	5	2.05	✓	Yes	
10	9	5	4	1.64	0	No	
11	19	15	4	1.64	0	No	
12	5	1	4	1.64	0	No	
13	13	9	4	1.64	0	No	
14	4	0	4	1.64	0	No	
15	6	2	4	1.64	0	No	Indeterminate
16	3	0	3	1.23	0	No	change
17	20	18	2	0.82	0	No	
18	9	10	–1	–0.41	0	No	
19	2	4	–2	–0.82	0	No	
20	7	10	–3	–1.23	0	No	
21	9	13	–4	–1.64	0	No	
22	3	7	–4	–1.64	0	No	
23	14	19	–5	–2.05	X	Yes	Reliably
24	10	20	–10	–4.10	X	Yes	deteriorated
25	18	33	–15	–6.15	X	Yes	
✓		= RC+, 0 = RC0, X = RC–, RC = reliable change.					

$C_{i \text{ (Standardized)}}$ or the RCI in determining reliable change; they yield consistent classifications. Indeed, both have been given the same label RC or RCI.[1] Table 1 displays a set of raw data to demonstrate that the same outcome results from the use of either classification method. Note, though, that $C_{i \text{ (Standardized)}}$ and RCI are conceptually different. $C_{i \text{ (Standardized)}}$ is a score given to an individual, and the RCI is a property of the measure. Below, I will refer to the equations above as defining the conventional JT method for RC/RCI (Evans *et al*., 1998; Maassen, 2000). Online calculators using the JT method are available to calculate RCI given relevant psychometric information (e.g. see https://www.psyctc.org/stats/rcsc1.htm).

Showing reliable change graphically

From the beginning (1984) Jacobson and colleagues used a particular kind of graph to illustrate change relative to the RCI and to reveal clinical significance for those cases achieving this (Evans *et al*., 1998). In this section I present a more elaborated form of this graph (Fig. 1), now called a modified Brinley plot (Blampied, 2017) and a second, simpler, dot-plot graph (Fig. 2) that displays RC, using the same data as $C_{i \text{ (Standardized)}}$. The graphs show four sets of data, which may be thought of either as the findings of an RCT comparing a control condition (Fig. 1A) with

---

[1]There is considerable variability in the literature on reliable change as to what is actually named RC and RCI. The standardized change score has been referred to using both labels. To distinguish the two calculations and for clarity and consistency I have used RCI to refer to the raw change score on the measure that indicates reliable change. When an individual's standardized change score meets the chosen criterion for reliable change I have used the label RC.
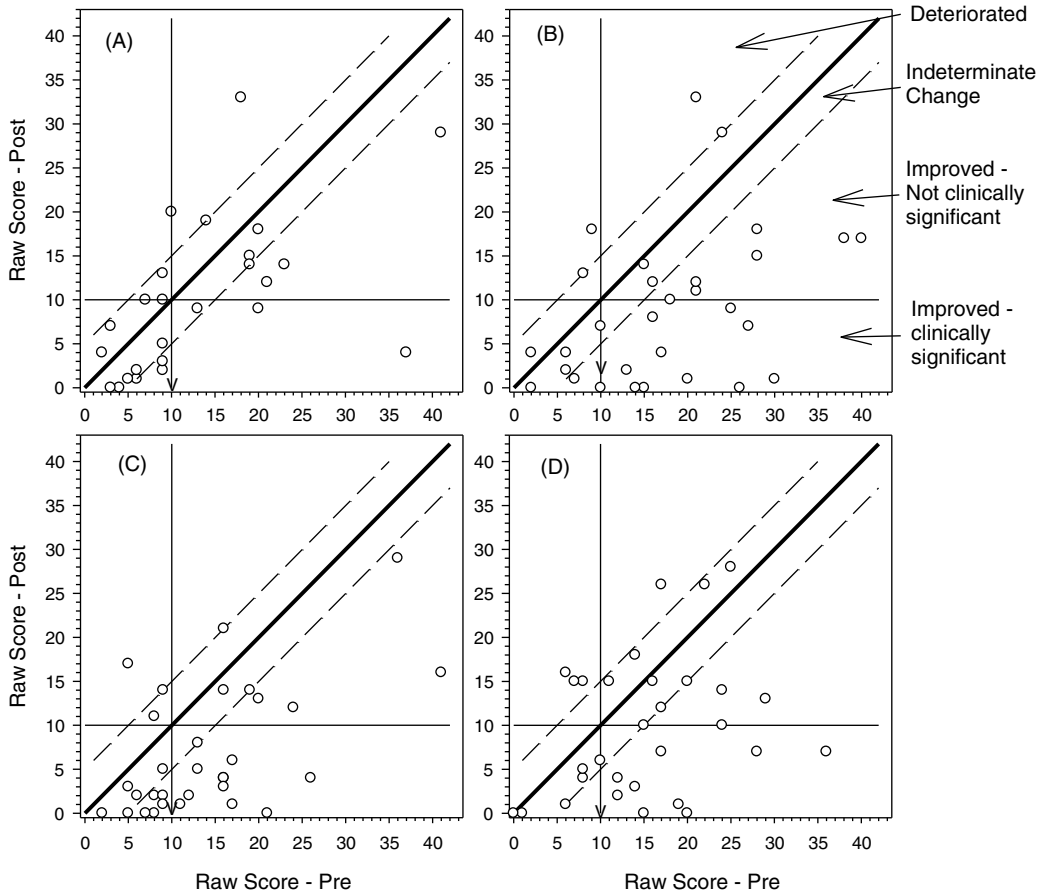
Figure 1. Modified Brinley plots of four sets of data that can be taken as coming from four arms of a randomized controlled trial or as representing four different clinical services. Each point represents the coordinate point for pre- and posttreatment raw scores. The continuous diagonal line represents the line of no change, the dashed diagonal lines represent the upper and lower boundaries of the reliable change index (set = 5) and the vertical and horizontal lines indicate a clinical cut-off score (set = 10). The arrowhead on the vertical line indicates the direction of change. The classification of cases falling in the different zones on the figure is shown to the right of panel B. Data are modified from Rucklidge *et al.* (2012) but for generality the specific scale has not been identified.

three treatment conditions (Fig. 1B, C, D), or a comparison of results achieved by four branches of a clinical service serving the same population, or possibly even four therapists/therapy teams within the same service (see below for more on how RC/RCI may be used in research and practice settings). For practitioners especially, once a graph template has been set up, data from subsequent cases can be added to track outcomes attained for cases treated over time.

A modified Brinley plot is a scatter plot where each individual's pre and post (or other t_1 vs t_2 data, such as post-treatment *vs* follow-up) is displayed as a coordinate point. Given that the x (t_1) and y (t_2) axis scales are the same, any individual whose scores are unchanged will have their data point lying on the 45 deg diagonal line of no change. If change has occurred, the data point will be shifted away from the diagonal to an extent proportional to the change. The distribution of scores pre and post can be seen by looking at the distribution relative to each axis (e.g. you can see in Fig. 1D that there are fewer cases in this condition who have pre scores below the clinical cut-off). Lines can be drawn parallel to the central diagonal to represent ±RCI. Cases lying between those
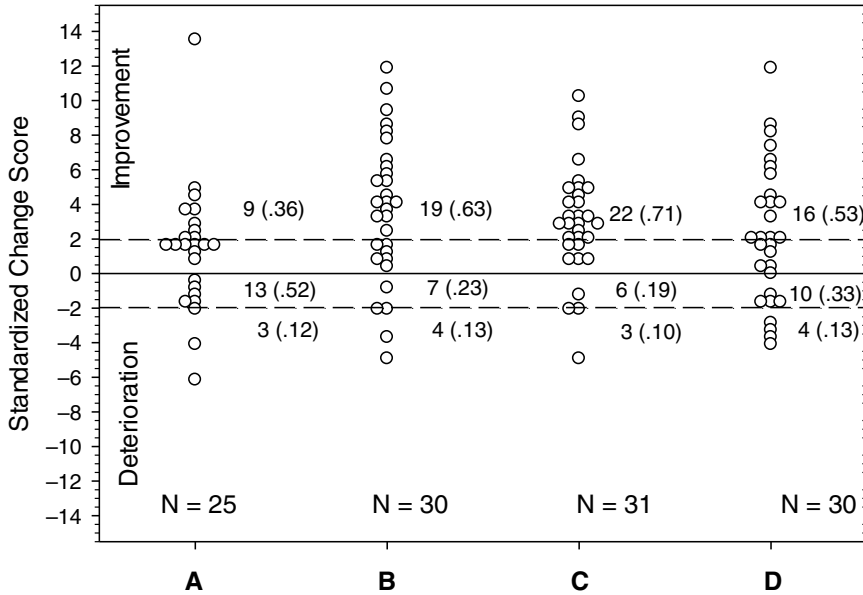
Figure 2. Dot plot scores of data from the same cases shown in Fig 1. Each data point represents the standardized change score (raw change score/standard error of the difference) of the case. Dashed lines show the boundaries for reliable positive change (improvement, +1.96) and deterioration (–1.96). The total number of cases in each set and the number (and proportion) in each category (Improved, Indeterminate, Deteriorated) is also shown.

limits are showing indeterminate change (i.e. insufficient to be deemed reliably changed). Depending on the direction of clinically desirable change on the measure (which can be shown as an arrow on the graph), cases whose data lie beyond the RCI limits are either improved or deteriorated, and those within the limits are indeterminate. Where horizontal and vertical lines are added to show clinical cut-offs, those who have improved and who also fall beyond the horizontal cut-off line can be classified as clinically significantly changed. Blampied (2017) has shown how other information, such as ES, means and confidence intervals, can be added to the plot to assist further interpretation (see also Black *et al.* 2019). Graphs of this kind (Figs 1 and 2) can display large amounts of data for visual inspection, revealing trends and outcomes much more immediately than data tables, and if different symbols are used for different aspects of cases or treatments, may reveal moderator variable effects and treatment by severity interactions (Follette and Callaghan, 2001; see Black *et al.*, 2019, and Rucklidge and Blampied, 2011, for examples).

Figure 2 shows the same data using a simpler dot plot. Each dataset (from Fig. 1A–D) is plotted on the same graph with separate groups of cases spaced across the *x*-axis. The *y*-axis displays $C_{i\,(\text{Standardized})}$. While in principle this is unbounded about zero, for any specific measure the maximum and minimum scores are given by the maximum possible difference score/S_{Diff}, so that for a measure with maximum range of, for example, 0–50 (maximum difference score = ±50) and $S_{\text{Diff}} = 5$, $C_{i\,(\text{Standardized})}$ could range from +10 to –10. The line of no change is shown as a horizontal line at zero, with ±1.96 boundaries as dashed horizontal lines above and below the zero line. Cases with data points above the upper dashed line are improved and those below the lower line are deteriorated, with those between being classified as indeterminate. Note that for these data larger change scores indicate positive change because reduced scores indicate clinical improvement on the measure. That would reverse if increased scores indicated improvement.

Research and professional uses of RC/RCI

Much of the literature on RC/RCI has been directed at its use in research, although Jacobson and colleagues clearly considered that practitioners as well as researchers would benefit from a clear specification of reliable change and clinical significance in treatment outcome studies. Nevertheless, beyond this common interest in clinical significance and its RC pre-requisite, somewhat different uses of RC/RCI can be discerned for researchers and professional practitioners.

In the case of a research study, classifying each participant's change as reliable is a further step that the investigator can take beyond reporting NHST statistics and effect sizes (both of which are nomothetic), providing a more complete, and some might argue, more honest, account of the findings. It also provides idiographic information in an otherwise entirely nomothetic context. Jacobson and colleagues argued that this should be routinely reported (Jacobson *et al.*, 1984; p. 349) and Lambert has echoed this (Lambert and Bailey, 2012; Lambert and Ogles, 2009). While such reporting adds important information about treatment outcome, Jacobson and colleagues (Jacobson *et al.*, 1984; Jacobson and Truax, 1991) and others since have warned that estimates of the efficacy of therapy are likely to be reduced by this practice. Jacobson and colleagues also argued that routine use of RC/RCI would motivate researchers to improve the quality of outcome measures, as, other things being equal, more reliable tests will have a smaller RCI and thus be more likely to find clinically significant change.

Furthermore, given that RC/RCI permits the classification of participants into a range of categories of response to treatment (Jacobson and Truax, 1991; Wise, 2004), this provides opportunities to explore the moderation and mediation of treatment effects and could then feed into further research on how to enhance the impact of treatment (Follette and Callaghan, 2001). This might profitably be extended into further analysis of what predicts maintenance of gains at and beyond follow-up.

Jacobson and colleagues also clearly saw that the proportion of cases in the study sample deemed RC+ (RC+%) constitutes an ES (Blampied, 2017) which complements ES measures such as Cohen's *d* (Lakens, 2013). Note that RC–% (the proportion who have deteriorated) may also be a useful measure; sometimes prevention of deterioration is an important therapy outcome (Aizik-Reebs *et al.*, 2021). The reporting of ES measures has become much more common than it was in 1984, with the standardized mean difference (Cohen's *d*) being popular (Ferguson, 2016). Abramowitz (1998) used an adapted *d* where the mean difference was standardized by S_{Diff} rather than the pooled *SD* to estimate the ES. This indicates if the change demonstrated by a 'typical' case – one whose scores were at the pre- and post-treatment means – would be considered reliably changed. Other researchers have since used this ES (e.g. Sheldrick *et al.*, 2001; Thompson *et al.*, 2021) in meta-analyses. Where studies incorporated into a meta-analysis report RC/RCI classifications this information can also be used to enrich meta-analytic findings (e.g. Cuijpers *et al.*, 2021).

The uses so far considered all apply to the analysis of data in hand after a study is completed. As Zarah and Hedge (2010) pointed out, RC/RCI also can be used prospectively in research. This could be done, for instance, in studies where some prior manipulation of one or more groups of participants is performed as part of the design (e.g. a mood manipulation) before all participants are then exposed to some primary treatment. Identifying, in advance of the application of the primary independent variable (IV), those who have demonstrated reliable change in response to the prior manipulation permits the researcher to strengthen their research in one of several ways. First, they might select for further participation in the study only those who have shown reliable change (e.g. whose mood has changed reliably in the appropriate direction), thereby strengthening the research by reducing one source of potential variance in the response to the IV. Alternatively, the researcher might classify participants into subgroups according to the RC+, RC– and RC0 classifications and examine the effect of this on the response to the primary IV (Zarah and Hedge, 2010).

For practitioners, understanding RC/RCI will clearly be useful insofar as this knowledge will assist them in the evaluation of research relevant to their practice. There may, however, be some more direct utility for the practitioner in using RC/RCI in routine practice because of its specifically idiographic nature (Jacobson *et al.*, 1984). It is a source of practice-based evidence (Barkham and Mellor-Clark, 2003) and can contribute to measurement-based care (Scott and Lewis, 2015). In making decisions about the course of therapy, practitioners can use RC on relevant measures to assess change in clients and thence the clinical significance of any change, although this would be only one element in a comprehensive clinical judgement of progress and outcome (de Sousa Costa and de Paula, 2015; Ojserkis *et al.*, 2014).

Other potential uses in practice for RC/RCI include its use to defend practice innovations to service management authorities. Demonstrating that an innovation produces reliable and clinically significant change in client outcomes may help practitioners confirm that the innovation is useful and possibly obtain support and resources for maintaining and extending the innovation (e.g. Draycott *et al.*, 2012). Also, the routine use of RC/RCI assessments may also contribute to professional development (Kelley, 2010) and to service audit processes through the identification of services, client categories, and diagnostic/functional domains where outcomes are, or are not, being reliably achieved (e.g. Deane *et al.*, 2013) and thus support constructive and remedial actions.

The use of RC/RCI in the audit of outcomes achieved by clinical services can be extended to include large-scale reviews at the regional, national, and potentially, international level. For example, the RCI is used as part of the Improving Access to Psychological Therapies programme in the United Kingdom (e.g. Barkham *et al.*, 2012; National Collaborating Centre for Mental Health, 2021). Important caveats apply to all such use of RC/RCI in audit and evaluation contexts. First, great care must be taken to ensure that the same algorithms and key psychometric coefficients are used in the RC/RCI computations (see below) to ensure that like is compared with like. Second, no decision regarding any particular therapy, therapist/therapy team, service or clinical system should be made exclusively on the basis of RC/RCI data. Such judgements should include wider considerations, including the clinical significance of the outcomes achieved (which is the primary reason for concern with RC in the first place).

Some issues and debates

Almost as soon as Jacobson and colleagues published their ideas others offered amendments, improvements, and replacements for the JT method. This identified a number of issues and engendered a variety of debates about the method, which continue to the present. In this section I will touch on several of these issues and debates that I see as particularly salient.

Calculating the SE_M in the JT method

Getting the SE_M for a measure is a key step. It requires two bits of data, s_1 and r_{xx} (equation (2) above). Jacobson and Truax (1991) were not very explicit or prescriptive in specifying the source of this information, but it is clear from their worked example that they envisaged it coming from an appropriate study reporting norms for the measure. This is fine in principle but not necessarily easy in practice. The SE_M is not a fixed property of a scale, like the number and wording of its items; it is a contextually determined estimate of the precision of the measure in a specific instance of measurement with a specific sample. It may be challenging for a contemporary investigator to find a norm study that matches her/his investigative context. For instance, while norms may be available for a scale for Western, anglophone participants, what norms should be used for participants from other linguistic, cultural, and ethnic groups, or, what effect might the passage of time have had on adolescent norms from a study done in the pre-internet era?

Careful judgement and defensible reasons are needed in the selection of data for the SE_M calculation (Kendall *et al.*, 1999).

Given that norm data have been found, there is not much debate about selecting s_1 – it is the *SD* of an appropriate norm group on the scale (but see McAleavey, 2021) – but the selection of the reliability coefficient has been more contentious. Classical test theory offers several reliability coefficients, of which internal consistency and test–retest reliability (r_{xx}) are the most relevant here. Jacobson and Truax (1991) recommended that r_{xx} should be used, but many subsequent investigators of RC have used internal consistency, particularly coefficient alpha (α; also known as Cronbach's α). Coefficient α can conveniently be estimated from a single test administration, and is widely reported in the psychometric literature. Its convenience and wide use has led to it being recommended (Lambert and Bailey, 2012) as standard practice for determining RC/RCI. Within the psychometric literature, however, coefficient alpha has been widely investigated and both critiqued and defended (Green and Yang, 2009; Raykov and Marcoulides, 2019) and alternatives, notably coefficient omega (ω), have been recommended (Flora, 2020). Investigators who have reason to prefer ω to α can substitute one for the other in their RC calculations, but I have not seen an example of this.

Recently, in a comprehensive critique of the classical JT method, McAleavey (2021) has argued that only test–retest reliability estimated over short inter-test intervals should be used, because it is the measure of reliability matching the pre–post repeated measures aspect of the data being analysed, and coefficient α should not be used. If implemented, this recommendation would preclude the use of RC/RCI for scales for which appropriate test–retest reliability could not be located, likely to lead to considerable restriction in the use of RC/RCI. Furthermore, test–retest reliability estimation is not a simple matter (Aldridge *et al.*, 2017; Weir, 2005) and even if a test–retest r_{xx} is located in the literature, it is not always clear that it is the right version. The commonly reported Pearson's r correlation is problematic because it is insensitive to systematic error across repeated measurements (Aldridge *et al.*, 2017).[2] For some measures (e.g. rating scales) the preferred coefficient for estimating test–retest reliability is the intraclass correlation coefficient (ICC), not Pearson's r (Aldridge *et al.*, 2017; Weir, 2005) but the calculation and interpretation of the ICC is complex. Once an appropriate ICC coefficient has been selected it can, and should, be used as r_{xx} in the calculation of SE_M (Weir, 2005). Given competing recommendations regarding the selection of a reliability coefficient and the strong influence reliability has on SE_M and the impact that in turn has on RC/RCI the field needs authoritative advice on the relative merits of the different reliability coefficients in the calculation of RC/RCI. This is a challenge that needs urgent attention by psychometricians. In the meantime, the selection and use of any particular reliability coefficient for the purposes of calculating RC/RCI needs careful specification and justification.

Accommodating practice effects

The scales commonly used to assess symptom severity in CBT, such as those used in research on the treatment of anxiety and depression, are typically assumed to be able to repeatedly measure the current psychological state of the individual without their score reflecting any substantive influence of them having previously taken the test, i.e. the scales are assumed not to exhibit practice effects. In contrast, tests of learning and of cognitive performance are known to be affected by practice effects, so that an individual's scores will change over

---

[2]Restricting the use of RC/RCI is consistent with the central argument of McAleavey (2021). At the time of writing this paper is available as an online pre-print, not yet formally published following peer review. It argues that RC/RCI should be used only when the circumstances of its use meet a set of quite restrictive criteria, of which the use only of test–retest reliability is one instance. It also demonstrates an alternative method for the calculation of RC/RCI and notes some alternative methods by which reliable change might be assessed.

repeated administrations even in the absence of any true change (Duff, 2012). This has led some, especially neuropsychologists, to suggest ways that the raw change score can be adjusted to take account of practice effects. There are a range of methods for adjusting the calculation of RC/RCI to take account of practice effects, ranging from adjusting each individual's raw change score by a constant estimate of the practice effect derived from the mean data of some appropriate control group to more complex regression-based adjustments (Duff, 2012). Some of these alternatives also involve changes to the calculation of SE_M. Readers who may need to take account of practice effects should consult Duff (2012) for more detail about available methods.

Alternative methods for calculating RC/RCI

Over time, a considerable number of alternative formulas and statistical methods have been suggested as improved ways of obtaining RC/RCI (for selected summaries, see Hynton-Bayre, 2010; Wise, 2004). Earlier versions of these alternatives have been evaluated by examining their mathematical adequacy and selectively compared with one another using both synthetic and real data (see Lambert and Bailey, 2012, for a review). No alternative to the classical JT method approach has received general endorsement, and Lambert and Bailey (2012) recommend the routine use and reporting of the JT method, in combination with other methods if the investigator desires. Others concur, often emphasizing that the JT method is relatively simple to use and produces results that are often similar to those from more complex methods (e.g. Atkins *et al.*, 2005; Bauer *et al.*, 2004; Ronk *et al.*, 2016; but for reservations, see McAleavey, 2021).

Recently, investigators have been reporting other, statistically advanced, ways of estimating SE_M and other parameters. For example, Hays *et al.* (2021) and Maydeu-Olivares (2021) have explored the use of item response theory (IRT) to calculate RC/RCI. IRT is an alternative to classical test theory for psychometric test development. Other examples have explored ways to extend the calculation of RC/RCI across multiple time points, using moderated, non-linear factor analysis to estimate SE_M (Morgan-Lopez *et al.*, 2022). As yet, these methods are too new to have received much evaluative attention, but for investigators with large data samples and the requisite technical expertise, these may become useful techniques. They are less likely ever to be of direct use to the practitioner. In the meantime, the classical JT method continues regularly to be used in evaluations of the outcome of CBT (e.g. Anastopoulos *et al.*, 2021; Aizik-Reebs *et al.*, 2021; Auyeung *et al.*, 2020; Mathews *et al.*, 2022; Timulack *et al.*, 2022).

Conclusion

Thirty-eight years after it was first proposed by Jacobson and colleagues, is the concept of RC and the JT method still useful to researchers and practitioners of CBT? If one accepts that clinical significance is of great importance and that determining if individual change is reliably larger than might be due to measurement error alone is a necessary pre-requisite to determining clinical significance, then the answer is certainly yes. This is still the case even for those who adopt a different definition of clinical significance than that proposed by Jacobson *et al.* (1984). Additionally, in a general sense, having a constant reminder of the ubiquity of measurement error is a good thing, not least because it should motivate us to improve our measures and to take care to select measures with high reliability. Also, the concept is a counterweight to the hegemony of the group statistical approach that so dominates psychology. It gives us an idiographic tool in an otherwise nomothetic tool box.

Over the years the JT method has been rigorously challenged, but it has survived, and can claim the virtue of simplicity and the recommendation of popularity relative to alternatives (Wise, 2004) and has received authoritative endorsement and recommendation for routine use (Lambert and Bailey, 2012; but see McAleavey, 2021 for reservations). Clearly, though, those using the JT

method need to be aware of some of its weaknesses. If an investigator suspects their measure is influenced by practice effects, one of the methods described by Duff (2012) should be used instead. Care must be taken to select the proper reliability estimate for the JT formula for SE_M. Unfortunately, doubts about the relative merits of internal consistency and test–retest reliability estimates remain unresolved and we lack expert guidance about this. Better ways of estimating key parameters and of incorporating multiple measures beyond just two time points may well be developed relatively soon, and would be welcome when they occur (McAleavey, 2021).

Practitioners may well find RC/RCI to be useful in a variety of ways that improve clinical practice, but it goes without saying, that no decision to amend, change or terminate therapy or to classify a client in some way (e.g. as fit for work) should be made on the basis of RC/ RCI alone. It is just one piece of information to feed into the matrix of clinical judgement. Also, when the context (e.g. degenerative disease, chronic severe conditions, developmental disabilities) makes the 1.96 criterion too severe and likely to find few if any clients reliably improved, then a more lenient criterion should be used.

In any use of RC/RCI in research or practice, it is essential that the method used be clearly and comprehensively described and choices among options justified. Details given should include the exact formulas used, the source of any norm data, and if local sample data are to be used instead, exactly how this is done. There has been, and I think we will have to accept that there will continue to be, a plurality of ways of estimating RC/RCI, just as there are a plurality of ways of testing for statistical significance, judging clinical significance, and estimating effect size. This has implications for the generality of any specific conclusions about reliable change that require careful consideration. Only if each method is carefully described when it is used will we be able to make ongoing judgements about the merits or otherwise of its use in specific instances and thereby improve our use of the method. With many aspects of our methodology we have to accept that there is no single mechanical ritual (Cohen, 1994) or statistical idol (Gigerenzer, 2018), recourse to which will absolve us of the necessity of judgement. That is clearly true for determining reliable change.

Key practice points

(1) Using reliable change (RC) or the reliable change index (RCI) for a specific measure permits the practitioner to know if the degree to which any specific case has changed during treatment is greater than might have occurred just due to measurement error alone.

(2) This is a source of practice-based evidence and can help determine if clinically significant change has occurred for any case.

(3) Evidence about achievement (or otherwise) of reliable change may assist with supporting innovations in practice and with audit and professional development processes.

Further reading

Barker, C., Pistrang, N., & Elliot, R. (2016). *Research Methods in Clinical Psychology* (3rd edn). Chichester, UK: Wiley.

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. New York, USA: Routledge.

Lambert, M. J., & Bailey, R. J. (2012). Measures of clinically significant change. In H. Cooper (Editor-in-Chief), *APA Handbook of Research Methods in Psychology: Vol 3. Data Analysis and Research Publication* (pp. 147–160). Washington, DC, USA: American Psychological Association.

References

Abramowitz, J. S. (1998). Does cognitive-behavioral therapy cure obsessive-compulsive disorder? A meta-analytic evaluation of clinical significance. *Behavior Therapy*, *29*, 339–335.

Aizik-Reebs, A., Soham, A., & Bernstein, A. (2021). First do no harm: an intensive experience sampling study of adverse effects to mindfulness training. *Behaviour Research & Therapy*, *145*. doi: 10.1016/j.brat.2021.103941

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test–retest reliability of psychological measures. *European Psychologist*, 22. doi: 10.1027/1016-9040/a000298

Anastopoulos, A. D., Langberg, J. M., Eddy, L. D., Silva, P. J., & Labban, J. D. (2021). A randomized controlled trial examining CBT for college students with ADHD. *Journal of Consulting & Clinical Psychology*, *89*, 21–33.

Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: does it matter what method we use? *Journal of Consulting & Clinical Psychology*, *73*, 982–989

Auyeung, K., Hawley, L., Grimm, K., & Rowa, K. (2020). Fear of negative evaluation and rapid response to treatment during cognitive behaviour therapy for social anxiety disorder. *Cognitive Therapy & Research*, *44*, 526–537.

Balluerka, N., Gomez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *European Journal of Research Methods for the Behavioural & Social Sciences*, *1*, 55–70.

Barker, C., Pistrang, N., & Elliot, R. (2016). *Research Methods in Clinical Psychology* (3rd edn). Chichester, UK: Wiley.

Barkham, M., & Mellor-Clark, J. (2003). Bridging evidence-based practice and practice-based evidence: developing a rigorous and relevant knowledge for the psychological therapies. *Clinical Psychology & Psychotherapy*, *10*, 319–327.

Barkham, M., Stiles, W. B., Connell, J., & Mellor-Clark, J. (2012). Psychological treatment outcomes in routine NHS services: what do we mean by treatment effectiveness? *Psychology and Psychotherapy: Theory, Research, and Practice*, *85*, 1–16.

Barlow, D. H. (ed). (2014). *The Oxford Handbook of Clinical Psychology*. Oxford, UK: Oxford University Press.

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The Scientist-Practitioner: Research and Accountability in Clinical and Educational Settings*. New York, USA: Pergamon.

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives in Psychological Science*, *4*, 19–21.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change* (3rd edn). Boston, MA, USA: Pearson Education.

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

Bergin, A. E., & Strupp, H. (1970). New directions in psychotherapy research. *Journal of Abnormal Psychology*, *76*, 15–26.

Bergin, A. E., & Strupp, H. (1972). *Changing Frontiers in the Science of Psychotherapy*. Chicago, IL, USA: Aldine.

Black, S. R., Blampied, N., Arnold, L. E., & Fristad, M. A. (2019). Is evidence-based treatment helping my patient? Utilizing modified Brinley plots to measure clinical change. *Clinical Psychology: Science and Practice*, *26*. doi: 10.1111/cpsp.12272

Blampied, N. M. (2017). Analysing therapeutic change using modified Brinley plots: history, construction, and interpretation. *Behavior Therapy*, *48*, 115–127.

Blampied, N. M. (2013). Single-case research and the scientist-practitioner ideal in applied psychology. In G. Madden (Editor-in-Chief), *Handbook of Behavior Analysis, Vol 1 Methods and Principles* (pp. 177–197). Washington, DC, USA: American Psychological Association.

Borenstein, M. (2012). Effect size estimation. In H. Cooper (Editor-in-Chief). *APA Handbook of Research Methods in Psychology: Vol 3. Data Analysis and Research Publication* (pp. 131–146). Washington, DC, USA: American Psychological Association.

Chambless, D. L. & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*, 7–18.

Chambless, D. L. & Hollon, S. D. (2012). Treatment validity for intervention studies. In H. Cooper (Editor-in-Chief). *APA Handbook of Research Methods in Psychology: Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological* (pp. 529–552). Washington, DC, USA: American Psychological Association.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single-subject: an alteration of the RC Index. *Behavior Therapy*, *17*, 305–308.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist*, *49*, 997–1003.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*. doi: 10.1098/rsos.140216

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied Behavior Analysis* (3rd edn). Upper Saddle River, NJ, USA: Pearson.

Cuijpers, P., Karyotaki, E., Chiarova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response remission, reliable change and deterioration: a meta-analysis. *Acta Psychiatrica Scandinavica*. doi: 10.1111/acps.13335

Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge, UK: Cambridge University Press.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting & Clinical Psychology*, *62*, 75–82.

de Sousa Costa, D., & de Paula, J. J. (2015). Usefulness of the reliable change index for psychology and psychiatry in clinical practice: a case report of cognitive-behavioral therapy. *Clinical Neuropsychiatry*, *12*, 135–138.

Deane, F. P., Kelly, P. J., Crowe, T. P., Coulson, J. C., & Lyons, C. B. (2013). Clinical and reliable change in an Australian residential substance use program using the Addiction Severity Index. *Journal of Addictive Diseases*, *32*, 194–205. https://doi.org/10.1080/10550887.2013.795470

Doyle, L., Brady, A.-M., & Bryne, G. (2009). An overview of mixed methods research. *Journal of Research in Nursing*, *14*, 175–185.

Draycott, S., Kirkpatrick, T., & Askari, R. (2012) An idiographic examination of patient progress in the treatment of dangerous and severe personality disorder: a reliable change index approach. *Journal of Forensic Psychiatry & Psychology*, *23*, 108–124

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*, 248–261.

Evans, C., Marginson, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental Health*, *1*, 70–72. https://ebmh.bmj.com/content/1/3/70

Ferguson, C. J. (2016). An effect size primer: a guide for clinicians and researchers. In A. E. Kazdin (ed), *Methodological Issues and Strategies in Clinical Research* (pp. 301–310). American Psychological Association. https://doi.org/10.1037/14805-020

Fidler, F., Cumming, G., Thomason, N., Panuzzo, D., Smith, J., Fyffe, P., . . . & Schmitt, R. (2005). Towards improved statistical reporting in the *Journal of Consulting & Clinical Psychology*. *Journal of Consulting & Clinical Psychology*, *73*, 136–143.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS*. doi: 10.1073/pnas.1711978115

Fisher, R. A. (1935). *The Design of Experiments*. London, UK: Oliver & Boyd.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods & Practices in Psychological Science*, *3*, 484–501.

Follette, W. C., & Callaghan, G. M. (2001). The evolution of clinical significance. *Clinical Psychology: Science & Practice*, *8*, 431–435.

Gigerenzer, G. (2018). Statistical rituals: the replication delusion and how we got there. *Advances in Methods & Practices in Psychological Science*, *1*, 198–218.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrica*, *74*, 121–135.

Hays, R. D., Spirtzer, K. L., & Reise, S. P. (2021). Using item response theory to identify responders to treatment: examples with the patient-reported outcomes measurement information system (PROMIS®) physical function scale and emotional distress composite. *Psychometrica*, *86*, 781–792.

Hubbard, R. (2004). Alphabet soup: blurring the distinctions between *p*'s and *a*'s in psychological research. *Theory & Psychology*, *14*, 295–327.

Hubbard, R., Parsa, R. A., & Luthy, M. R. (1987). The spread of statistical testing in psychology. *Theory & Psychology*, *7*, 545–554.

Hynton-Bayre, A. D. (2010). Deriving reliable change statistics from test-retest normative data: comparison of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*, 244–256.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting & Clinical Psychology*, *59*, 12–19.

Kelley, P. J. (2010). Calculating clinically significant change: applications of the Clinical Global Impressions (CGI) scale to evaluate client outcomes in private practice. *Clinical Psychologist*, *14*, 107–111.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 285.

Kent, D., & Hayward, R. (2007). When averages hide individual differences in clinical trials. *American Scientist*, *95*, 60–68.

Kiesler, D. J. (1971). Experimental designs in psychotherapy research. In A. E. Bergin & S. L. Garfield (eds), *Handbook of Psychotherapy and Behavior Change* (pp. 36–74). London, UK: Wiley.

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Millbank Quarterly*, 82, 661–687.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, article 863. doi: 10.3389/fpsyg.2013.00863

Lambdin, C. (2012). Significance tests as sorcery: science is empirical – significance tests are not. *Theory & Psychology*, 22, 67–90. doi: 10.1177/0959354311429854

Lambert, M. J., & Bailey, R. J. (2012). Measures of clinically significant change. In H. Cooper (Editor-in-Chief), *APA Handbook of Research Methods in Psychology: Vol 3. Data Analysis and Research Publication* (pp. 147–160). Washington, DC, USA: American Psychological Association. doi: 10.1037/13621-007

Lambert, M. J., & Ogles, B. M. (2009). Using clinical significance in psychotherapy outcome research: the need for a common procedure and validity data. *Psychotherapy Research*, 39, 493–501.

Lamiell, J. T. (1998). 'Nomothetic' and 'Idiographic': contrasting Windelband's understanding with contemporary usage. *Theory & Psychology*, 8, 23–38.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suarez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: the APA Publications and Communications Board task force report. *American Psychologist*, 73, 26–46.

Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical & Experimental Neuropsychology*, 22, 622–632.

Mathews, S. R., Elizabeth, M., Roberts, L. N., & Nixon, R. D. V. (2022). Client vs clinicians' standards of clinically meaningful change and the effects of treatment expectations on therapeutic outcomes in individuals with posttraumatic stress disorder. *Behavior Therapy*. doi: 10.1016/j.beth.2021.12.007

Maydeu-Olivares, A. (2021). Assessing the accuracy of errors of measurement. Implications for assessing reliable change in clinical settings. *Psychometrica*, 86, 793–799.

McAleavey, A. A. (2021). When (not) to rely on the reliable change index. https://osf.io/download/619b4998dbcf80493eda59d

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: applications for health research. *Qualitative Health Research*, 8, 362–376.

Morgan-Lopez, A. A., Saavedra, L. M., Ramirez, D. D., Smith, L. M., & Yaros, C. A. (2022). Adapting the multi-level model for estimation of the reliable change index (RCI) with multiple time points and multiple sources of error. *International Journal of Methods in Psychiatric Research*. https://doi.org/10.1002/mpr.1906

National Collaborating Centre for Mental Health (2021). *The Improving Access to Psychological Therapies Manual (Version 5)*. https://www.england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual-v5.pdf

Nickerson, J. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.

Norcross, J. C., VandenBos, G. R., & Freedheim, D. K. (Editors-in-Chief) (2016). *APA Handbook of Clinical Psychology*. Washington, DC, USA: American Psychological Association.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2021). Clinical significance: History, application and current practice. *Clinical Psychology Review*, 21(3), 421–446.

Ojserkis, R., Morris, B., & McKay, D. (2014). Pediatric obsessive-compulsive disorder: an illustration of intensive family-based treatment delivered via a web camera. *Clinical Case Studies*, 3, 68–79.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you. *Educational & Psychological Measurement*, 79, 200–210.

Ronk, F. R., Hooke, G. R., & Page, A. C. (2016). Validity of clinically significant change classifications yielded by Jacobson-Truax and Hageman-Arrindell methods. *BMC Psychiatry*, 16, 1–9.

Rucklidge, J. J., Andridge, R., Gorman, B., Blampied, N. M., Gordon, H., & Boggis, A. (2012). Shaken but unstirred? Effects of micronutrients on stress and trauma after an earthquake: RCT evidence comparing formulas and doses. *Human Psychopharmacology Clinical & Experimental*, 27, 440–454.

Rucklidge, J. J., & Blampied, N. M. (2011). Post-earthquake psychological functioning in adults with attention-deficit/hyperactivity disorder: positive effects of micronutrients on resilience. *New Zealand Journal of Psychology*, 40, 51–57.

Schneider, J. W. (2015). Null hypothesis significance tests: a mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102, 411–432.

Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive & Behavioral Practice*, 22, 49–59.

Sheldrick, R. C., Kendall, P. C., & Heimberg, R. G. (2001). The clinical significance of treatments: a comparison of three treatments for conduct disordered children. *Clinical Psychology: Science & Practice*, 8, 418–430.

Speelman, C. P., & McGann, M. (2016). Editorial: Challenges to mean-based analysis in psychology: the contrast between individual people and general science. *Frontiers in Psychology*, 7, article 1234. doi: 10.3389/fpsyg.2016.01234

Spring, B. (2007). Evidence-based practice in clinical psychology: what it is, why it matters, what you need to know. *Journal of Clinical Psychology*, *63*, 613–631.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge, MA, USA: Harvard University Press.

Thompson, E. M., Destree, L., Albertella, L., & Fontenelle, L. F. (2021). Internet-based acceptance and commitment therapy: a transdiagnostic systematic review and meta-analysis for mental health outcomes. *Behavior Therapy*, *52*, 492–507.

Timulack, L., Keogh, D., Chigwedere, C., Wilson, C., Ward, F., Hevey, D., Griffin, P., Jacobs, L., Hughes, S., Vaughan, C., Beckham, K., & Mahon, S. (2022). A comparison of emotion-focused therapy and cognitive-behavioral therapy in the treatment of generalized anxiety disorder: results of a feasibility randomized controlled trial. *Psychotherapy*, *59*, 84–95.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., & Thompson, B. (2000). Reporting practices and editorial policies regarding statistical significance and effect size. *Theory & Psychology*, *10*, 413–425.

Valsiner, J. (1986). *The Individual Subject and Scientific Psychology*. New York, USA: Plenum.

Wasserstein, R. L., & Lazar, N. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*, 129–133. doi: 120.1080/00031305.2016.1154108

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength & Conditioning Research*, *19*, 231–240.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, *54*, 594–604.

Windelband, W. (1894/1998). History and natural science (translated from the German by J. T. Lamiell). *Theory & Psychology*, *8*, 1–22.

Wise, E. A. (2004). Methods for analysing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment*, *82*, 50–59.

Wright, A. G. C., & Hallquist, M. N. (eds) (2020). *The Cambridge Handbook of Research Methods in Clinical Psychology*. Cambridge, UK: Cambridge University Press.

Wright, D. B. (2009). Ten statisticians and their impacts for psychologists. *Perspectives on Psychological Science*, *4*, 587–597.

Zarah, D., & Hedge, C. (2010). The reliable change index: why isn't it more popular in academic psychology? *PsyPag Quarterly*, *76*, 14–19.