International Journal of Social Research Methodology

ISSN: 1364-5579 (Print) 1464-5300 (Online) Journal homepage: www.tandfonline.com/journals/tsrm20

Improving the reliability of the Reliable Change Index

Alexander O. Crenshaw & Candice M. Monson

To cite this article: Alexander O. Crenshaw & Candice M. Monson (07 Nov 2025): Improving the reliability of the Reliable Change Index, International Journal of Social Research Methodology, DOI: 10.1080/13645579.2025.2585286

To link to this article: https://doi.org/10.1080/13645579.2025.2585286

INTERNATIONAL JOURNAL OF SOCIAL RESEARCH METHODOLOGY
https://doi.org/10.1080/13645579.2025.2585286

Routledge
Taylor & Francis Group

RESEARCH ARTICLE

Improving the reliability of the Reliable Change Index

Alexander O. Crenshaw [a] and Candice M. Monson [b]

[a]Department of Psychological Science, Kennesaw State University, Kennesaw, GA, USA; [b]Department of Psychology, Toronto Metropolitan University, Toronto, ON, Canada

ABSTRACT

The reliable change index (RCI) is a tool for evaluating change at the individual level. It compliments standard group-level change estimates and is central to evaluating clinical significance for interventions. In principle, the RCI provides common criteria for evaluating individual change. However, current practices use sample-specific estimates to create these criteria. Because sample estimates are subject to sampling error, these criteria are also subject to sampling error and therefore differ across studies. We illustrate how current practices can lead to differing criteria for reliable change and use simulations to identify the impact of sampling error on the RCI. Excessive error in the RCI began for the average sample when $N < 30$, and samples only comfortably avoided the risk of excessive error when $N > 100$. Finally, small errors in estimating a measure's reliability sometimes had profound effects on the RCI. Recommendations on use of the RCI are provided.

ARTICLE HISTORY
Received 17 January 2025
Accepted 30 October 2025

KEYWORDS
Reliable Change Index; clinical significance; small samples; sampling error; evaluating change

Studies evaluating change over time commonly make use of the Reliable Change Index (RCI; Jacobson & Truax, 1991). Whereas standard evaluation of change focuses on the average change of a group, the RCI quantifies the number of *individuals* within a group who changed beyond that expected by chance. The RCI compliments group-averaged estimates and is particularly widely used when evaluating interventions (Kraemer et al., 2006; Lambert & Ogles, 2014). Moreover, like a ruler establishes a common criterion for measuring length, the RCI aims to establish common criteria for evaluating change across different studies and outcomes (Lambert & Ogles, 2009, 2014).

However, current practice commonly involves using sample-specific estimates to create the criteria upon which reliable change is evaluated (Crenshaw et al., 2025). In other words, what constitutes the amount of change required to be considered reliable is often based on estimates from a given sample. Since sample estimates are subject to sampling error, these criteria are subject to sampling error and therefore differ across studies, even those involving the same outcome and population. When the RCI is over- or underestimated in a study due to sampling error, it systematically biases the evaluation of reliable change for all individuals in the study. We call this phenomenon the 'variable ruler problem,' in which the criteria – or rulers – for evaluating reliable change differ across studies for reasons other than genuine differences across settings. In this paper, we

CONTACT Alexander O. Crenshaw acrenshaw@torontomu.ca Department of Psychological Science, Kennesaw State University, 402 Bartow Ave NW, Kennesaw, GA, 30144, USA

describe how common practices for computing reliable change can lead to 'variable rulers' when evaluating change, present the results of simulations using real and generated data demonstrating the impact of sampling error on the variable ruler problem, and make recommendations for future studies that use the RCI.

Importance of evaluating reliable change

Standard statistical evaluation of change over time involves estimating average change at the group level, such as change between two or more timepoints in an overall sample (e.g. in a longitudinal study) or between-group differences in change (e.g. in a randomized experiment). A well-documented limitation of these approaches is that they obscure information about the experiences of individuals within those groups (Atkins et al., 2005; Kraemer et al., 2006; Lambert & Ogles, 2009, Lambert and Ogles, 2014). Individual outcomes can look very different and still have the same group-level outcomes. For example, an intervention that moderately helps all individuals might have the same average outcome as one that is exceptionally helpful to some and of no benefit to others. Interventions may also result in consistent but trivial benefits that result in statistically significant differences at the group level, but without any practical benefit to participants (Kraemer et al., 2006).

The rise of reliable change came about in response to these limitations of group average estimates (Jacobson & Truax, 1991) and is in line with idiographic approaches more broadly (e.g. Fisher et al., 2018). Reliable change evaluates, for each individual, whether they experienced enough change to rule out chance variation due to measurement error. It aligns with the concept of *detected difference* in Langkaas et al.'s (2018) summary of five types of change to evaluate in clinical interventions and is the mantlepiece for evaluating clinical significance, which is recommended standard practice when evaluating clinical interventions (Kraemer et al., 2006; Lambert & Ogles, 2014). In addition to its ubiquity in clinical research and practice, reliable change is used across the social sciences. Some examples include studies investigating change in loneliness (Masi et al., 2011), compassion (Krieger et al., 2019), and parenting practices (Sanders et al., 2000); the effect of educational interventions (Mandy et al., 2016); and numerous studies investigating change in personality (e.g. Damian et al., 2019; Robins et al., 2001). Outside of the social sciences, reliable change has also been applied in the evaluation of cognitive decline (Chang et al., 2009), to assess change in health-related quality of life (Crosby et al., 2003), to validate a measure for the assessment of standing balance (Clark et al., 2010), and assessment of change in Asthma symptoms (Schatz et al., 2009).

Much has been written about the concept of reliable change (for reviews, see Atkins et al., 2005; Blampied, 2022; De Smet, 2024; Lambert & Ogles, 2009, 2014; McAleavey, 2024), and numerous methods have been proposed to operationalize it, including those based on classical test theory (e.g. Jacobson & Truax, 1991), item response theory (e.g. Hays et al., 2021), and latent variable modeling (e.g. Saavedra et al., 2021). However, as has been noted by many observers (Atkins et al., 2005; Blampied, 2022; Lambert & Ogles, 2014; McAleavey, 2024), Jacobson and Truax's (1991), Reliable Change Index is overwhelmingly the most commonly used. McAleavey (2024) noted that Jacobson and Truax (1991) was the most cited article in the history of the *Journal of Consulting and Clinical Psychology* 30 years after its

publication. Additionally, a recent methodological review of clinical trials published in clinical psychology journals from 2020 to 2023 found that 100% (23 of 23) of reviewed studies that evaluated reliable change and reported the method used Jacobson and Truax (1991) (Crenshaw et al., 2025). It is thus this method on which we focus in this paper, but we will return to the issue of alternative operationalizations in the Discussion.

Computation of the Reliable Change Index

Following terminology in Blampied (2022) and McAleavey (2024), Jacobson and Truax's (1991), Reliable Change Index (hereafter referred to as the JT method) is a combination of three equations:

$$Standard\ error\ of\ measurement(s_E) = s\sqrt{1 - r_{xx}} \tag{1}$$

$$Standard\ error\ of\ the\ difference\left(s_{diff}\right) = \sqrt{2 * {s_E}^2} \tag{2}$$

$$Reliable\ Change\ Index = 1.96 * s_{diff} \tag{3}$$

where s is the standard deviation of the outcome measure, r_{xx} is a measure of reliability, and the RCI is the amount of change in the measure needed to be considered reliable. Jacobson and Truax (1991) do not explicitly state which values of s or r_{xx} should be used. However, their example suggests s should be the standard deviation of either a control group, normal population, or pretreatment group. Similarly, their example uses test–retest reliability for r_{xx}, which is also advocated by McAleavey (2024), but this is an area of active disagreement, with some suggesting internal consistency instead (e.g. Lambert & Ogles, 2009, 2014).

Functionally, the RCI equation does several things. Equation (1) computes the expected variability in the measure due to measurement error. Equation (2) computes the expected variability when estimating change resulting from using two measurement time points that are each impacted by measurement error, reflecting the expected variability in estimates of change if no change had occurred. Finally, Equation (3) uses null-hypothesis significance testing logic to identify the threshold of change needed to rule out change due to normal variability alone, where $\alpha = .05$ (two-tailed). This is the threshold used to evaluate all individuals in a sample: those with change greater than the threshold are considered to have reliably improved (or deteriorated), and those with change less than the threshold have not. An alternative approach, as originally described in Jacobson and Truax (1991) and termed RC by Blampied (2022), and which is simply a rearrangement of terms from the RCI, is to compute a change score standardized on s_{diff} for each individual: $RC = \frac{Change}{s_{diff}}$. The RC for each person is then compared against the critical value for the chosen α (e.g. 1.96 for a two-tailed $\alpha = .05$).

The RCI equations can be combined to

$$Reliable\ Change\ Index = 1.96\sqrt{2\left[s\sqrt{1 - r_{xx}}\right]^2} \tag{4}$$

and then, to highlight the separate contributions of s and r_{xx} to the overall RCI, simplified to

$$Reliable\ Change\ Index = 1.96 * s * \sqrt{2 - 2r_{xx}} \tag{5}$$

The Reliable Change Index is Influenced by Sampling Error in *s* and r_{xx}

The RCI formula is thus made up of a variability term (s) and a reliability term ($\sqrt{2 - 2r_{xx}}$), plus a constant (1.96). The desired values to use in this formula are the population values for the standard deviation and measure reliability, σ and ρ_{xx}, respectively. Because population values are rarely known, the sample values, s and r_{xx}, must be used instead. For some instruments (e.g. Lambert et al., 2004; Marx et al., 2022), a standard RCI has been developed a priori and can be used by other studies using that same instrument, enabling reliable change to be evaluated according to consistent criteria across studies (see the Discussion for further consideration of this approach). As noted by Blampied (2022), this may have been the approach originally envisioned by Jacobson and Truax (1991). However, such premade RCI values are unavailable for many instruments (Blampied, 2022; Lambert & Ogles, 2014), and computation of RCI using s and r_{xx} from local data is extraordinarily common. For example, the previously-mentioned review of published clinical trials in psychology found that 85% of studies using the RCI computed it using sample estimates (Crenshaw et al., 2025).

As sample values are influenced by sampling error, error in these values causes error in the RCI. Whereas standard statistical tests account for sampling error through the standard error, the RCI ignores sampling error by only using the point estimates s and r_{xx} (this point was noted as far back as Jacobson and Truax (1991), but has received little attention since). Sampling error can also impede comparison across studies. When one reports, for example, that 50% of patients in a treatment sample showed reliable improvement, one presumably wants that 50% to be comparable to 50% from another study. However, given that the RCI is subject to sampling error, some studies will overestimate it and others will underestimate it. Because the RCI is the criterion against which change for all individuals in a study is evaluated, over- or underestimation of the RCI systematically biases the evaluation of reliable change for all individuals in the study.

The impact of sampling error on the RCI is currently unknown, so the comparability of reliable change results across studies is unclear. We set out to estimate the impact of sampling error in s and r_{xx} on the RCI, focusing on two empirical questions: 1) What is the individual and combined impact of variability in s and r_{xx} on the RCI? 2) What is the real-world impact of sampling error in s and r_{xx} on the RCI? In the discussion, we address whether these results warrant changing practices in how s and r_{xx} are used to compute the RCI.

Method

The overall goal of the study is to assess the extent to which sampling error impacts sample estimates of the RCI and its components (the variability term, s, and the reliability term, $\sqrt{2 - 2r_{xx}}$). Doing so requires emulating the collection of samples and assessing the extent to which the variability term, reliability term, and the RCI vary across such

samples. We approach this goal in three steps. First, we used Monte Carlo simulations that sample *s* at fixed values of r_{xx} to understand the extent to which specific magnitudes of error in r_{xx} estimation impact the RCI. As this approach is artificial from a sampling perspective, we next simulate samples for four hypothetical self-report questionnaires from a generated hypothetical population. Finally, to understand the impact of sampling error in real-world uses, our third approach applies resampling techniques using eight questionnaires from two real datasets.

In all cases, we use internal consistency as the operationalization of r_{xx} because it is widely used and can be easily computed from any sample, whereas test–retest reliability can only be computed with particular study designs (Blampied, 2022). Procedures were approved by the Research Ethics Board at Toronto Metropolitan University (#2023-136). Data and analysis code in R (R Core Team, 2022) are available online (https://doi.org/10.17605/OSF.IO/Z2KDU).

Data Selection/Generation

Sampling s at fixed values of r_{xx}

To obtain *s*, we randomly drew samples of size *N* (where $N = 10$ to 200) from the normal (*z*) distribution ($M = 0$, $SD = 1$), with 5,000 iterations per sample size. For r_{xx}, we explored a range of fixed values for the sample r_{xx} and the population equivalent, ρ_{xx}. Values ranged from .70 to .95, as .70 is a commonly used lower end estimate of acceptable internal consistency and .95 is the highest internal consistency typically observed for measures in the social sciences. Differences between r_{xx} and ρ_{xx} were limited to ≤.15.

Sampling both s and r_{xx} using simulated data

We first generated a population of 100,000 individuals by randomly generating 'true' scores on a hypothetical construct from the normal (*z*) distribution ($M = 0$, $SD = 1$). We created individual item scores for four hypothetical self-report questionnaires assessing the construct, where the questionnaires had either high ($\alpha = .95$) or low ($\alpha = .70$) internal consistency and either more (10) or fewer (4) items in the scale (resulting in four scale properties: high reliability/more items, high reliability/fewer items, low reliability/more items, and low reliability/fewer items). To create the scales, we generated scores on hypothetical items by adding random noise to the true score from the distribution $N(0, \sigma^2)$, where σ^2 was adjusted until the scale of 4 or 10 items had the desired value for α. The noise value, σ^2, for each item in a scale was identical to ensure equal item loadings on the construct. Next, we randomly selected samples of size *N* (*N*s = 10 to 200) from this population with replacement, with 5,000 iterations each.

Sampling both s and r_{xx} using real data

We used public data from two projects posted in online repositories. Data sources were selected with the goal of obtaining data from several different clinically relevant self-report questionnaires, with a range of values for internal consistency and for the number of items in the questionnaire. We searched for papers with open data badges published in the journal *Clinical Psychological Science* from 2022 through March of 2023 with large sample sizes ($N > 300$) and that publicly posted item-level data from commonly used

questionnaires. Ultimately, data from Tutorial 2 in Conway et al. (2022) and from Luke et al. (2022) met criteria.

Data from the Conway study (Tutorial 2) were a subset of respondents ($N = 725$) who reported being in a romantic relationship from a larger survey study of romantic experiences and mental health in an Australian community sample. We selected four outcome measures that are commonly used, simple to calculate, showed some heterogeneity in the internal consistency (although all were high), and with different numbers of items. Selected measures were the Patient Health Questionnaire (PHQ; a measure of depressive symptoms; Kroenke et al., 2001), the Brief Measure for Assessing Generalized Anxiety Disorder (GAD; Spitzer et al., 2006), the Obsessive-Compulsive Inventory-Revised (Foa et al., 2002), and the six-item Modified Quality of Marriage Index (QMI; Nazarinia et al., 2009). Internal consistency (Cronbach's α) was .89, .91, .91, and .95, respectively.

Luke et al. (2022) was a study about psychopathy and judgments of moral dilemmas in 443 individuals. We used data from the Self-Report Psychopathy Scale-Short Form (Paulhus et al., 2017), a measure of psychopathy with four subscales: affective, antisocial, interpersonal, and lifestyle. Subscales were examined separately. Internal consistency (α) was .66, .71, .80, and .77, respectively. These datasets can be found by accessing the original papers or through links at https://doi.org/10.17605/OSF.IO/Z2KDU. To mimic random sampling from a population, each full dataset was treated as a hypothetical population. From this population, we randomly selected samples of size N (Ns = 10 to 200) with replacement, with 5,000 iterations each.

Determining the impact of sampling error

For each step, we computed the variability term (s), the reliability term ($\sqrt{2 - 2r_{xx}}$), and the full RCI equation (Equation 5) in each sample. Then, for each sample, we computed the percent difference between the sample value and the respective population value ($\frac{Samplevalue - populationvalue}{Populationvalue} * 100$) separately for the variability term, reliability term, and full RCI. Next, we computed the mean absolute deviation of sample estimates from the population value across all 5,000 iterations. This value represents the average percent deviation (i.e. error) of sample estimates compared with population values for each of the measures. Average deviations above 10% were considered to be excessive error (Flora & Curran, 2004).[1] Finally, the mean only describes the typical deviation and may not well represent an individual study. We therefore also examined distributions of these deviations to identify when studies show concerning risk for excessive deviations from the population RCI, setting the threshold of risk at 20% or greater.

Results

Sampling s at fixed values of r_{xx}

Results for all 225 combinations of sample size and r_{xx} values are available online (https://doi.org/10.17605/OSF.IO/Z2KDU). Table 1 highlights a subset of those results for four sample sizes – small pilot studies ($N = 10$ and $N = 20$), the upper range of typical sample sizes for psychological intervention studies ($N = 200$), and a middle value ($N = 50$). The

Table 1. Individual and combined impact of variability in s and r_{xx} on the reliable change Index.

			Mean absolute deviation of the Reliable Change Index		
Population ρ_{xx}	Sample r_{xx}	$r_{xx} - \rho_{xx}$ difference	Variability term (s)	Reliability term ($\sqrt{2-2r_{xx}}$)	Full equation
$N = 10$					
Any	$r_{xx} = \rho_{xx}$	None	**19%**	0%	**19%**
.95	.90	−.05	**19%**	**41%**	**41%**
.95	.80	−.15	**19%**	**100%**	**94%**
.85	.90	.05	**19%**	**18%**	**24%**
.80	.70	−.10	**19%**	**22%**	**27%**
.70	.85	.15	**19%**	**29%**	**32%**
.70	.75	.05	**19%**	9%	**20%**
$N = 20$					
Any	$r_{xx} = \rho_{xx}$	None	**13%**	0%	**13%**
.95	.90	−.05	**13%**	**41%**	**40%**
.95	.80	−.15	**13%**	**100%**	**98%**
.85	.90	.05	**13%**	**18%**	**20%**
.80	.70	−.10	**13%**	**22%**	**23%**
.70	.85	.15	**13%**	**29%**	**30%**
.70	.75	.05	**13%**	9%	**15%**
$N = 50$					
Any	$r_{xx} = \rho_{xx}$	None	8%	0%	8%
.95	.90	−.05	8%	**41%**	**41%**
.95	.80	−.15	8%	**100%**	**99%**
.85	.90	.05	8%	**18%**	**19%**
.80	.70	−.10	8%	**22%**	**22%**
.70	.85	.15	8%	**29%**	**30%**
.70	.75	.05	8%	9%	**11%**
$N = 200$					
Any	$r_{xx} = \rho_{xx}$	None	4%	0%	4%
.95	.90	.05	4%	**41%**	**41%**
.95	.80	.15	4%	**100%**	**100%**
.85	.90	−.05	4%	**18%**	**18%**
.80	.70	.10	4%	**22%**	**22%**
.70	.85	−.15	4%	**29%**	**29%**
.70	.75	−.05	4%	9%	9%

Note: ρ_{xx} and r_{xx} = population and sample internal consistency, respectively; s = sample standard deviation of measure. Numbers crossing 10% error threshold are bolded.

three rightmost columns show mean absolute deviations of sample estimates from population values for the variability term (s), reliability term ($\sqrt{2-2r_{xx}}$), and the full RCI equation (Equation (5)). Bolded values in these columns exceed the 10% error threshold.

Unsurprisingly, for the variability term, the average deviation increased as sample size decreased. When $N \leq 30$, average deviations exceeded the 10% threshold of excessive error. For the reliability term, any discrepancy between the sample and population reliability was highly impactful, and all but one average deviation (when $\rho_{xx} = .70$ and $r_{xx} = .75$) exceeded the 10% threshold. Of note, differences at higher reliabilities were much more impactful than at lower reliabilities: a .05 difference between .95 and .90 led to a 41% average deviation in the reliability term, whereas a .05 difference between .70 and .75 led to a 9% average deviation. For the full RCI, the only instances in which the average deviation of the RCI was below 10% was (when $N > 40$ and $r_{xx} = \rho_{xx}$). Together, these results show that sample estimates of the RCI and each of its components have excessive

Table 2. Impact of sampling error in s and r_{xx} on the Reliable Change Index: simulated data.

			Mean absolute deviation of the Reliable Change Index		
Variable	Population ρ_{xx}	Number of items	Variability term (*s*)	Reliability term ($\sqrt{2-2r_{xx}}$)	Full equation
N = 10					
High r_{xx}, more items	.95	10	**19%**	**23%**	6%
High r_{xx}, fewer items	.95	4	**19%**	**25%**	**11%**
Low r_{xx}, more items	.70	10	**19%**	**23%**	6%
Low r_{xx}, fewer items	.70	4	**19%**	**25%**	**11%**
N = 20					
High r_{xx}, more items	.95	10	**13%**	**14%**	4%
High r_{xx}, fewer items	.95	4	**13%**	**16%**	8%
Low r_{xx}, more items	.70	10	**13%**	**14%**	4%
Low r_{xx}, fewer items	.70	4	**13%**	**16%**	8%
N = 50					
High r_{xx}, more items	.95	10	8%	9%	3%
High r_{xx}, fewer items	.95	4	8%	**10%**	5%
Low r_{xx}, more items	.70	10	8%	9%	3%
Low r_{xx}, fewer items	.70	4	8%	**10%**	5%
N = 200					
High r_{xx}, more items	.95	10	4%	4%	1%
High r_{xx}, fewer items	.95	4	4%	5%	2%
Low r_{xx}, more items	.70	10	4%	4%	1%
Low r_{xx}, fewer items	.70	4	4%	4%	2%

Note: ρ_{xx} and r_{xx} = population and sample internal consistency; *s* = sample standard deviation of measure. Numbers crossing the 10% error threshold are bolded.

error at small sample sizes and when the reliability term is estimated with any known amount of error.

Sampling both s and r_{xx} using simulated and real data

Mean deviation

Table 2 shows results from the simulated data, and Table 3 shows results for each measure from Conway et al. (2022, tutorial 2) and Luke et al. (2022). Tables again show sample sizes of $N = 10$, 20, 50, and 200, and the OSF page contains the full tables with all sample sizes. Average deviations, of sample estimates from the population value, for the variability term of the RCI equation were consistently below 10% at $N \geq 40$ for simulated data and at $N \geq 100$ for Conway and Luke. Average deviations for the reliability term were consistently below 10% at $N \geq 75$ for simulated data and at $N \geq 100$ for

Table 3. Impact of sampling error in s and r_{xx} on the Reliability Change Index: Conway and Luke Studies.

			Mean absolute deviation of the Reliable Change Index		
Variable	Population ρ_{xx}	Number of items	Variability term (*s*)	Reliability term ($\sqrt{2-2r_{xx}}$)	Full equation
N = 10					
PHQ	.89	9	**21%**	**26%**	**11%**
GAD	.91	7	**20%**	**25%**	**12%**
OCI	.91	18	**31%**	**34%**	**14%**
QMI	.95	6	**24%**	**31%**	**21%**
Affective Psychopathy	.66	7	**19%**	**24%**	**10%**
Antisocial Psychopathy	.61	8	**26%**	**29%**	**17%**
Interpersonal Psychopathy	.80	7	**20%**	**25%**	9%
Lifestyle Psychopathy	.77	7	**19%**	**24%**	**10%**
N = 20					
PHQ	.89	9	**15%**	**17%**	8%
GAD	.91	7	**13%**	**16%**	8%
OCI	.91	18	**21%**	**22%**	**10%**
QMI	.95	6	**16%**	**20%**	**15%**
Affective Psychopathy	.66	7	**13%**	**15%**	7%
Antisocial Psychopathy	.61	8	**18%**	**18%**	**11%**
Interpersonal Psychopathy	.80	7	**13%**	**16%**	6%
Lifestyle Psychopathy	.77	7	**13%**	**16%**	7%
N = 50					
PHQ	.89	9	9%	**10%**	5%
GAD	.91	7	8%	9%	5%
OCI	.91	18	**13%**	**13%**	6%
QMI	.95	6	**10%**	**12%**	9%
Affective Psychopathy	.66	7	8%	9%	4%
Antisocial Psychopathy	.61	8	**11%**	**11%**	7%
Interpersonal Psychopathy	.80	7	8%	**10%**	4%
Lifestyle Psychopathy	.77	7	8%	**10%**	4%
N = 200					
PHQ	.89	9	4%	4%	2%
GAD	.91	7	3%	4%	2%
OCI	.91	18	6%	6%	3%
QMI	.95	6	4%	5%	4%
Affective Psychopathy	.66	7	4%	4%	2%
Antisocial Psychopathy	.61	8	5%	5%	3%
Interpersonal Psychopathy	.80	7	4%	5%	2%
Lifestyle Psychopathy	.77	7	4%	5%	2%

Note: In each section, the top four outcome measures are from Conway et al. (2022, tutorial 2) and the bottom four are subscales of the Self-Report Psychopathy Scale-Short Form (Paulhus et al., 2017) from Luke et al. (2022). ρ_{xx} and r_{xx} = population and sample internal consistency; *s* = sample standard deviation of measure. Numbers crossing the 10% error threshold are bolded. RCI = Reliable Change Index, Psy. = Psychopathy. PHQ = Patient Health Questionnaire (Kroenke et al., 2001), GAD = Brief Measure for Assessing Generalized Anxiety Disorder (Spitzer et al., 2006), OCI = Obsessive-Compulsive Inventory-Revised (Foa et al., 2002), QMI = Modified Quality of Marriage Index (Nazarinia et al., 2009).

Conway and Luke. Average deviations in the full equation (rightmost column of Tables 2 and 3) varied considerably by the specific measure but were consistently below 10% at $N \geq 20$ in simulated data, at $N \geq 50$ in Conway, and at $N \geq 30$ in Luke. When excluding the QMI, which had the worst deviations, average deviations were consistently below 10% when $N \geq 30$ across samples.

Notably, average deviations in the full equation were considerably smaller than those for each term alone, suggesting sampling error in *s* and r_{xx} may offset one another.

Indeed, Pearson correlations between the estimates of s and r_{xx} from individual samples were positive and large, ranging from $r = .62$ to .94 across outcomes and samples. Consequently, when the variability term of the equation is higher, the reliability term tends to be lower ($\sqrt{2 - 2r_{xx}}$ decreases as r_{xx} increases). As a result, sampling error in each term partially cancels out the other, resulting in lower average deviation in the full equation.

As indicated by the higher average deviations, sampling error in r_{xx} had a greater impact than sampling error in s. This pattern was more pronounced when using simulated data and less pronounced when using real data. This pattern is consistent with, but smaller than, results from the previous section. Also consistent with results in the previous section, there was some evidence that sampling error in r_{xx} was more impactful at higher values of r_{xx} in the real data: measures with higher r_{xx} values tended to have greater mean deviation in the reliability term of the equation than those with lower r_{xx} values. However, the reliability of the outcome measure had no impact on the average deviation in the simulated data. We did not observe a clear pattern as to whether the number of items in a scale impacted mean deviation in the RCI in the real data, but did for the simulated data. In simulated data, scales with four items had double or nearly double the average deviation of scales with ten items on the full RCI.

Distribution of deviations across samples

As the mean deviation may not well represent what happens in a single study, we also examined distributions of deviations for the full RCI. Figures 1 and Figure 2 show results across different percentiles of the samples for each outcome in the simulated data and in Conway, respectively (results for Luke were similar to Conway and can be found on the OSF page). The y-axis represents the percent absolute deviation in the RCI, and the x-axis represents the percent of samples that have a deviation as large or larger than the y-axis value. Text labels are provided for the worst 10% and 20% of samples, and for the median sample.

These results provide context to the mean deviation values. The 10% error threshold was crossed in 20% or more samples when N was between 30 (PHQ) to as high as 100 (QMI). Errors were smaller for the simulated data, crossing the error threshold in 20% or more of samples when N was between 10 and 30. These results show that, although average sample deviations mostly stayed below the 10% error threshold when $N > 30$, the probability of a given sample having excessive error in the RCI may be uncomfortably high until sample sizes are 100 or more.

Discussion

The RCI provides an important complement to standard group-based evaluation of change and is a widely used metric for evaluating change in the social sciences, particularly in clinical interventions. However, the common practice of using sample estimates to create the criteria against which each individual's change is evaluated – the RCI – can systematically bias its evaluation in a given sample and create differing criteria for reliable change across studies. This paper explored the impact of sampling error on estimates of the RCI toward determining if a change in practice is warranted.

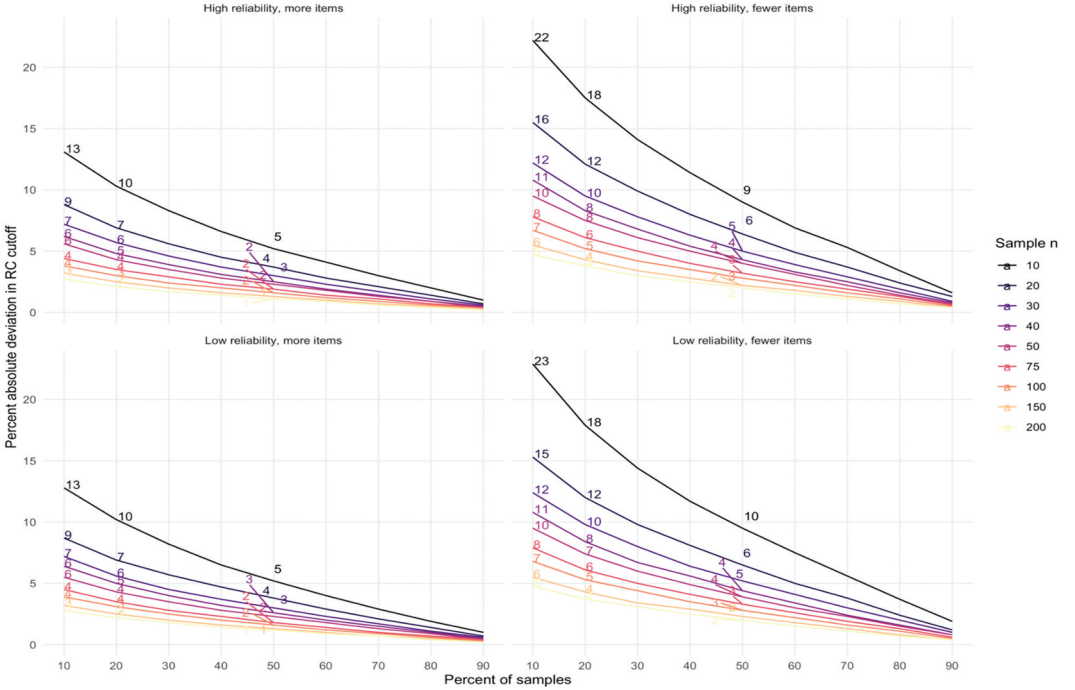

Figure 1. Percent of samples with [y] or greater deviation from the population value for the reliable change Index: simulated data. The y-axis represents the percent absolute deviation in the Reliable Change Index, and the x-axis represents the percent of samples that have a deviation as large or larger than the y-axis value. Text labels are provided for the worst 10% and 20% of samples, and for the median sample (50%). Example to aid interpretation: for the high reliability/more items measure (top left panel) when N = 10, 10% of samples (x-axis) have a deviation from the population value of 13% (y-axis) or greater, 20% of samples have a deviation of 10% or greater, and 50% of the samples have a deviation of 5% or greater.

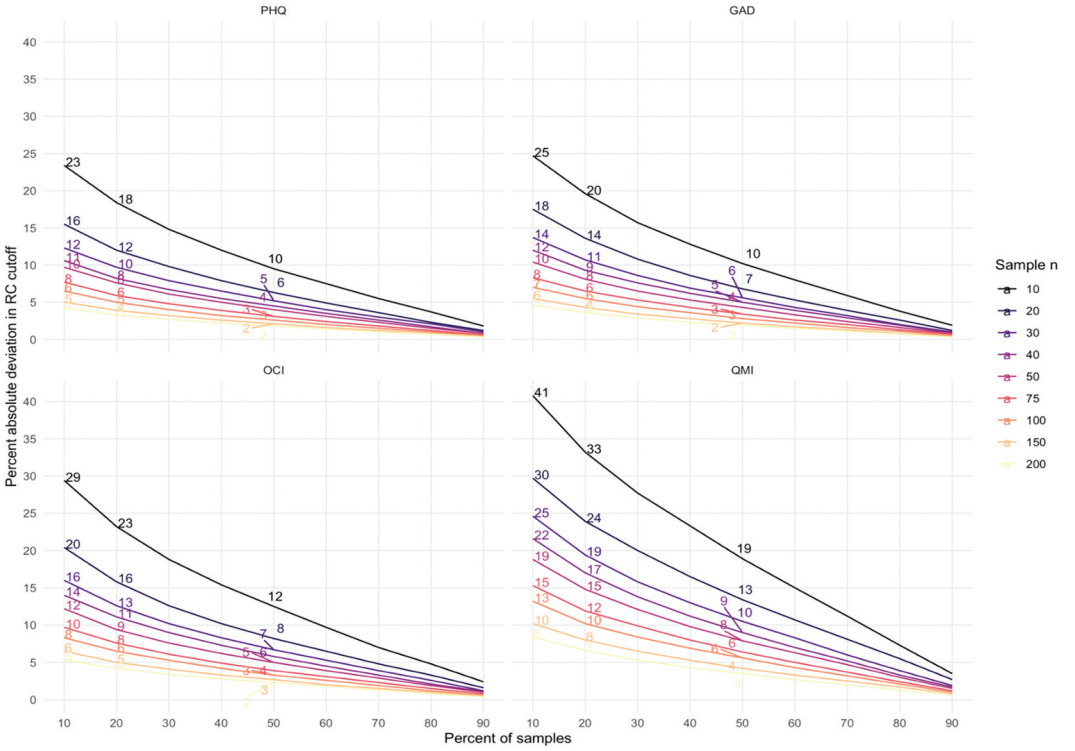

Figure 2. Percent of samples with [y] or greater deviation from the population value for the reliable change Index: Conway. The y-axis represents the percent absolute deviation in the Reliable Change Index, and the x-axis represents the percent of samples that have a deviation as large or larger than the y-axis value. Text labels are provided for the worst 10% and 20% of samples, and for the median sample (50%). Example to aid interpretation: for the PHQ (top left panel) when *N* = 10, 10% of samples (x-axis) have a deviation from the population value of 23% (y-axis) or greater, 20% of samples have a deviation of 18% or greater, and 50% of the samples have a deviation of 10% or greater. PHQ = Patient Health Questionnaire, GAD = Brief Measure for Assessing Generalized Anxiety Disorder, OCI = Obsessive-Compulsive Inventory-Revised, QMI = Modified Quality of Marriage Index.

When sampling s at fixed values for r_{xx} and ρ_{xx}, we found sampling error was larger for smaller sample sizes, and nearly any discrepancy between the sample and population r_{xx} resulted in excessive error regardless of sample size, up to doubling (or halving) of the population value. While artificial from a sampling perspective, this approach is informative when the population value of r_{xx} is either known or was previously estimated with high precision. Using a more realistic sampling process with generated and real data, we found sampling error in each component of the equation offset the other such that average deviations were considerably smaller in the full RCI compared with in the variability term or reliability term alone. Ultimately, for all but one outcome measure, average deviation in the full RCI stayed below the threshold for sample sizes of $N \geq 30$.

Instead of average error, a researcher planning a study may be more concerned with the chances of having excessive error in the particular study at hand. For simulated data, the sample size required to ensure the risk of excessive error stays below 20% for a given study ranged from $N > 10$ in the best case scenario to $N > 30$ in the worst case. For real data, required sample sizes ranged from $N > 20$ to $N > 100$.

Summary of factors impacting sampling error for the Reliable Change Index

Unsurprisingly, sample size was the most important factor impacting sampling error in the RCI. Average absolute deviations were small to negligible in the largest samples, and the chance of excessive error in a single study was consistently below 20% when sample size was above 100. In contrast, average deviations were often excessive when $N < 30$, which are typical sample sizes for pilot studies for psychological interventions. This pattern is noteworthy because the RCI is widely believed to be especially useful in small samples, when standard statistical tests are underpowered (Lambert & Ogles, 2009). However, these results show that the RCI is poorly estimated when N is small, potentially biasing evaluation of reliable change in these samples.

Finally, when the error in the r_{xx} estimate is known (e.g. if sampling from the same population as another, much larger, study), average deviation from the population value ranges from 9% to as high as 100%. Outside of these cases, error in r_{xx} was somewhat more impactful on the cut-off than error in s with real data, but not with simulated data. Finally, measures with fewer items had greater sampling error in the RCI compared with those with more items.

Additional sources of variability

The present investigation focused on cases in which researchers use the same formula for the RCI and use the same operational definitions for s and r_{xx}. However, as previously mentioned, numerous other methods have been proposed to evaluate reliable change (see Atkins et al., 2005; Blampied, 2022; De Smet, 2024; McAleavey, 2024). Moreover, even if studies all use the Jacobson and Truax (1991) method, its terms are underspecified in the original formula. The worked example identifies s as one of three possible standard deviations: of a control group, normal population, or pretreatment experimental group. Additionally, their example uses test–retest reliability for r_{xx}, but use of this operationalization is not explicitly instructed, and whether to use test–retest reliability or internal consistency for r_{xx} is an area of active debate (Blampied, 2022; Lambert & Ogles, 2014;

McAleavey, 2024). McAleavey (2024) argues that test–retest reliability with a short retest window is ideal because it matches the pre-post nature of the data, but Blampied (2022) note that computing test–retest has its own pitfalls for this purpose, and that this requirement would prevent computation of the RCI for many measures given a lack of existing test–retest estimate. We echo Blampied (2022) in calling for authoritative advice by psychometricians toward resolving this disagreement. Ultimately, use of different formulas for the RCI or different operationalization of s and r_{xx} will magnify the differences observed in the present investigations. Consequently, real world comparability of the RCI across studies is likely worse than shown in this study.

Is a change in practice warranted?

There are some advantages to the common practice of using individual sample estimates of s and r_{xx} to compute the RCI. This method is simple and convenient, and it is already standard practice in empirical research to compute standard deviations and internal consistency (if used for r_{xx}). Additionally, variability and reliability for a measure can differ between populations, settings, and usage, so sample values have better ecological validity by ensuring the estimate is relevant for population being studied. However, the results from the present study suggest that some change in practices is warranted.

One option, implemented for some widely used measures (e.g. Lambert et al., 2004; Marx et al., 2022), is to use a premade RCI for a given measure, developed using a large and representative sample such as that in a validation study, to be used by all studies involving that measure in the same or a similar population. Major advantages of this approach include ensuring the RCI value is minimally impacted by sampling error due to the large sample, is adequately representative of the population, and that different studies use equivalent criteria for evaluating reliable change so results are comparable across studies. The disadvantages are a loss in ecological validity – when there is a mismatch between the population or context for which the RCI was established and the population-(s) or context for which it is used – and a lack of available premade RCIs for many instruments. Perhaps due to its simplicity, intuitiveness, historical factors, and past evidence of equivalent performance to more complicated methods at the time (e.g. Atkins et al., 2005), the JT method has thrived in usage despite the availability of methods with better construct validity for many uses (McAleavey, 2024). Readers are encouraged to consider alternative methods that best suit the purpose for a given study. However, it is likely the JT method is here to stay and, when used, it may be most useful in providing a standard criterion across studies against which reliable change is evaluated. Consequently, we recommend using a premade RCI if one exists and was developed in a comparable population.

When a premade RCI not available, the best option largely depends on sample size. At $N > 100$, using one's own sample values of s and r_{xx} is appropriate and low risk. In contrast, given that average deviation for multiple measures started crossing the 10% threshold when $N < 30$, these studies should use estimates of s and r_{xx} from elsewhere if available. Measure validation studies are ideal as they often involve large and representative samples, and having a default choice maximizes comparability of reliable change results across studies. For sample sizes between 30 and 100, researchers should use their

judgment based on the study sample size and availability and representativeness of thresholds from a larger sample.

When a premade RCI is not available, we considered a 'mix and match' approach (e.g. Shalom et al., 2020) of using s from one's own sample in combination with the best available estimate of r_{xx} from another study. This approach would minimize sampling error in r_{xx}, which has a disproportionate effect on the RCI, while closely matching the population when estimating s. However, results showed that error is minimized when using estimates of s and r_{xx} from the same sample. Consequently, the mix and match approach is not recommended.

Recommendations for computing the reliable change Index

(1) Use a premade RCI threshold if one has been previously developed from a large sample of a comparable population.
(2) If a premade RCI threshold is not available or there is a substantial population mismatch between the premade threshold and one's study,
   (a) If $N > 100$, use the sample estimates of s and r_{xx} to compute the RCI.
   (b) If $N < 30$, use estimates of s and r_{xx} from a large-N study that best matches one's population of study. Measure validation studies are a good default choice.
   (c) If $30 \leq N \leq 100$, make a determination based on the study sample size and availability and representativeness of thresholds from a larger sample.
(3) Do not use estimates of s and r_{xx} from different samples.
(4) Justify the choice of method for the RCI and its specific operationalization, and discuss limitations and biases that may result. Document such choices ahead of time in a study preregistration due to the impact that these choices may have on the strength of apparent evidence for change in a study.

Limitations and future directions

Several limitations should be mentioned. First, we explored a limited set of outcome measures. Second, reliable change is just one aspect of the larger construct known as clinical significance (Jacobson & Truax, 1991). Third, we focused on the Jacobson and Truax (1991) operationalization and, as mentioned previously, numerous other methods have been proposed. Many such methods were developed in response to criticisms of the JT method, including its limitation to considering just two time points (McAleavey, 2024; Morgan-Lopez et al., 2022), inattention to other sources of error or how measurement error can change over time (e.g. Saavedra et al., 2021), use of nomothetic estimates for an idiographic analysis (e.g. McAleavey, 2024), and that categorizations from this method may be either overly conservative (McAleavey, 2024) or overly liberal (e.g. Morgan-Lopez et al., 2022). The JT method remains the dominant method for evaluating reliable change and thus was the focus of the present study, but other methods may be more appropriate for a given study.

Unfortunately, use of alternative methods that may be more suitable for a given study comes at the cost of muddying the comparability of reliable change results across studies.

Given the plethora of methods available, the variability in the RCI in actual use will be larger than that observed here when focused on just one method. Several studies have sought to quantify such differences resulting from use of different methods (e.g. Atkins et al., 2005; Ferrer & Pardo, 2014; Hays et al., 2021; Morgan-Lopez et al., 2022; Saavedra et al., 2021) and specific operationalizations of the same method (e.g. Crenshaw et al., 2024; Crenshaw et al., 2025), but more work is needed.

Conclusion

The RCI is an important complement to standard group-based estimates in the evaluation of change. We showed that common practices for computing the RCI renders it vulnerable to bias due to sampling error, particularly at small samples. We recommend a change in practice when computing the RCI, away from the current default of always using estimates of s and r_{xx} from one's sample and to a more principled consideration based on sample size. Recommendations include using a premade RCI threshold developed from a comparable population when available. If such value is not available, we recommend using a population-matched larger-N study when $N < 30$, using sample values when $N > 100$, and making a case-by-case determination for sample sizes between 30 and 100.

Note

1. The 10% threshold is commonly used for bias, not variability, in parameter estimates. Although we apply it here to a measure of variability in the estimate, not bias per se, usage in this context is akin to bias in the traditional usage of the threshold: when the RCI is over- or underestimated due to sampling error, it systematically biases the evaluation of reliable change for all individuals in the study.

Author contributions

CRediT: **Alexander O. Crenshaw:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing; **Candice M. Monson:** Funding acquisition, Resources, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded in part by the Canadian Institutes of Health Research Foundation Grant [139193] and Operating Grant [450266], and from the Canadian Institute for Military and Veterans Health Research True Patriot Love Research Initiative 6023235-985815. The grant sponsor played no role in study design; the collection, analysis, and interpretation of data; the writing of this paper; or the decision to submit this paper for publication.

Notes on contributors

Alexander O. Crenshaw, PhD, is an Assistant Professor of Psychology at Kennesaw State University and Director of the Couples and Methodology Lab. His research improves the design, analysis, and interpretation of clinical trials in psychology, in addition to treatment evaluation and the science of intimate relationship functioning. He has served as the quantitative expert on multiple federally funded clinical trials. His work centers open science practices and appears in leading journals across clinical and relationship science.

Candice M. Monson, PhD, is Professor of Psychology at the Toronto Metropolitan University. She is a leading expert on traumatic stress and the use of individual and conjoint therapies to treat PTSD. She has been funded by the U.S. Department of Veterans' Affairs, National Institute of Mental Health, Centers for Disease Control and Prevention, Department of Defense, and Canadian Institutes of Health for her research on interpersonal factors in traumatization and individual- and conjoint-based interventions for PTSD. She has published over 185 peer-reviewed publications, 8 books, and 42 book chapters on the development, evaluation, and dissemination of PTSD treatments.

ORCID

Alexander O. Crenshaw http://orcid.org/0000-0002-7453-2398
Candice M. Monson http://orcid.org/0000-0001-6179-0788

Data availability statement

Data and code are available at the Open Science Framework (https://doi.org/10.17605/OSF.IO/Z2KDU).

References

Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting & Clinical Psychology*, *73*(5), 982–989. https://doi.org/10.1037/0022-006X.73.5.982

Blampied, N. M. (2022). Reliable change and the reliable change index: Still useful after all these years? *The Cognitive Behaviour Therapist*, *15*, e50. https://doi.org/10.1017/S1754470X22000484

Chang, E. L., Wefel, J. S., Hess, K. R., Allen, P. K., Lang, F. F., Kornguth, D. G., Arbuckle, R. B., Swint, J. M., Shiu, A. S., Maor, M. H., & Meyers, C. A. (2009). Neurocognition in patients with brain metastases treated with radiosurgery or radiosurgery plus whole-brain irradiation: A randomised controlled trial. *The Lancet Oncology*, *10*(11), 1037–1044. https://doi.org/10.1016/S1470-2045(09)70263-3

Clark, R. A., Bryant, A. L., Pua, Y., McCrory, P., Bennell, K., & Hunt, M. (2010). Validity and reliability of the Nintendo Wii Balance board for assessment of standing balance. *Gait & Posture*, *31*(3), 307–310. https://doi.org/10.1016/j.gaitpost.2009.11.012

Conway, C. C., Forbes, M. K., & South, S. C. (2022). A hierarchical taxonomy of psychopathology (HiTOP) primer for mental health researchers. *Clinical Psychological Science*, *10*(2), 236–258. https://doi.org/10.1177/21677026211017834

Crenshaw, A. O., MacDonald, A., Carr, K., Shull, A., Pope, S., & Blackmon, M. (2025). Inconsistent application of the reliable change index in clinical psychology trials leads to noisy estimates of individual change. [preprint]. https://doi.org/10.31234/osf.io/d9amr_v1

Crenshaw, A.O., Narine, A., Hashi, M., Carr, K., Derangula, T., & Woo, Y. (2024). Variability in operationalizing the standardized mean difference and reliable change in clinical trials. In A. Crenshaw & A. De Nadai (Eds.), Novel Methods in CBT Research and Practice [Symposium].

Presented at the annual convention of the Association for Behavioral and Cognitive Therapies (ABCT), Philadelphia, PA.

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, *56*(5), 395–407. https://doi.org/10.1016/S0895-4356(03)00044-1

Damian, R. I., Spengler, M., Sutu, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality & Social Psychology*, *117*(3), 674–695. https://doi.org/10.1037/pspp0000210

De Smet, M. M. (2024). Why (not) rely on patients' perspectives? Time for a meaningful standard to measure and define patient-specific change in clinical psychology. *Clinical Psychology Science & Practice*, *31*(3), 370–374. https://doi.org/10.1037/cps0000213

Ferrer, R., & Pardo, A. (2014). Clinically meaningful change: False positives in the estimation of individual change. *Psychological Assessment*, *26*(2), 370–383. https://doi.org/10.1037/a0035419

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Hays, R. D., Spritzer, 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. *Psychometrika*, *86*(3), 781–792. https://doi.org/10.1007/s11336-021-09774-1

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting & Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: Clinical, research, and policy importance. *Jama*, *296*(10), 1286–1289. https://doi.org/10.1001/jama.296.10.1286

Krieger, T., Reber, F., von Glutz, B., Urech, A., Moser, C. T., Schulz, A., & Berger, T. (2019). An internet-based compassion-focused intervention for increased self-criticism: A randomized controlled trial. *Behavior Therapy*, *50*(2), 430–445. https://doi.org/10.1016/j.beth.2018.08.003

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lambert, M. J., Morton, J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R., & Burlingame, G. (2004). *Administration and scoring manual for the outcome questionnaire (OQ-45.2)* (3rd ed.). American Professional Credentialing Services LLC.

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*, *19*(4/5), 493–501. https://doi.org/10.1080/10503300902849483

Lambert, M., & Ogles, B. M. (2014). Using clinical significance in psychotherapy outcome research. In W. Lutz & S. Knox (Eds.), *Quantitative and qualitative methods in psychotherapy research* (1st ed., 378–406). Routledge. https://doi.org/10.4324/9780203386071

Langkaas, T. F., Wampold, B. E., & Hoffart, A. (2018). Five types of clinical difference to monitor in practice. *Psychotherapy*, *55*(3), 241–254. https://doi.org/10.1037/pst0000194

Luke, D. M., Neumann, C. S., & Gawronski, B. (2022). Psychopathy and moral-dilemma judgment: An analysis using the four-factor model of psychopathy and the CNI model of moral decision-making. *Clinical Psychological Science*, *10*(3), 553–569. https://doi.org/10.1177/21677026211043862

Mandy, W., Murin, M., Baykaner, O., Staunton, S., Cobb, R., Hellriegel, J., & Skuse, D. (2016). Easing the transition to secondary education for children with autism spectrum disorder: An evaluation of the systemic transition in education programme for autism spectrum disorder (STEP-ASD). *Autism*, *20*(5), 580–590. https://doi.org/10.1177/1362361315598892

Marx, B. P., Lee, D. J., Norman, S. B., Bovin, M. J., Sloan, D. M., Weathers, F. W., Keane, T. M., & Schnurr, P. P. (2022). Reliable and clinically significant change in the clinician-administered PTSD scale for DSM-5 and PTSD checklist for DSM-5 among male veterans. *Psychological Assessment*, *34*(2), 197–203. https://doi.org/10.1037/pas0001098

Masi, C. M., Chen, H.-Y., Hawkley, L. C., & Cacioppo, J. T. (2011). A meta-analysis of interventions to reduce loneliness. *Personality and Social Psychology Review*, *15*(3), 219–266. https://doi.org/10.1177/1088868310377394

McAleavey, A. A. (2024). When (not) to rely on the reliable change index: A critical appraisal and alternatives to consider in clinical psychology. *Clinical Psychology Science & Practice*, *31*(3), 351–366. https://doi.org/10.1037/cps0000203

Morgan-Lopez, A. A., Saavedra, L. M., Ramirez, D. D., Smith, L. M., & Yaros, A. C. (2022). Adapting the multilevel model for estimation of the reliable change index (RCI) with multiple timepoints and multiple sources of error. *International Journal of Methods in Psychiatric Research*, *31*(2), e1906. https://doi.org/10.1002/mpr.1906

Nazarinia, R. R., Schumm, W. R., & White, J. M. (2009). Dimensionality and reliability of a modified version of Norton's 1983 quality marriage index among expectant and new Canadian mothers. *Psychological Reports*, *104*(2), 379–387. https://doi.org/10.2466/PR0.104.2.379-387

Paulhus, D. L., Neumann, C. S., & Hare, R. D. (2017). Manual of the Hare self-report psychopathy scale. Multi-Health Systems.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, *69*(4), 617–640. https://doi.org/10.1111/1467-6494.694157

Saavedra, L. M., Morgan-López, A. A., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Fitzpatrick, S., & Lopez-Castro, T. (2021). Putting the patient back in clinical significance: Moderated nonlinear factor analysis for estimating clinically significant change in treatment for posttraumatic stress disorder. *Journal of Traumatic Stress*, *34*(2), 454–466. https://doi.org/10.1002/jts.22624

Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The triple p-positive parenting program: A comparison of enhanced, standard, and self-directed behavioral family intervention for parents of children with early onset conduct problems. *Journal of Consulting & Clinical Psychology*, *68*(4), 624–640. https://doi.org/10.1037/0022-006X.68.4.624

Schatz, M., Kosinski, M., Yarlas, A. S., Hanlon, J., Watson, M. E., & Jhingran, P. (2009). The minimally important difference of the asthma control test. *The Journal of Allergy and Clinical Immunology*, *124*(4), 719–723. https://doi.org/10.1016/j.jaci.2009.06.053

Shalom, J. G., Strauss, A. Y., Huppert, J. D., Andersson, G., Agassi, O. D., & Aderka, I. M. (2020). Predicting sudden gains before treatment begins: An examination of pretreatment intraindividual variability in symptoms. *Journal of Consulting & Clinical Psychology*, *88*(9), 809–817. https://doi.org/10.1037/ccp0000587

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092