Journal of Clinical and Experimental Neuropsychology
Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/ncen20

Principles of Defining Reliable Change Indices

Gerard H. Maassen

Published online: 09 Aug 2010.

To cite this article: Gerard H. Maassen (2000) Principles of Defining Reliable Change Indices, Journal of Clinical and Experimental Neuropsychology, 22:5, 622-632

To link to this article: http://dx.doi.org/10.1076/1380-3395(200010)22:5;1-9;FT622

Journal of Clinical and Experimental Neuropsychology
2000, Vol. 22, No. 5, pp. 622-632

1380-3395/00/2205-622$15.00
© Swets & Zeitlinger

Principles of Defining Reliable Change Indices

Gerard H. Maassen
Utrecht University, Faculty of Social Sciences, Department of Methodology and Statistics,
Utrecht, The Netherlands

ABSTRACT

In this article, several salient measures for determining reliable change are scrutinized. The classic null hypothesis method is compared with more recent procedures based on interval estimation of the true change, including Kelley's formula. The latter category of methods are shown to entail serious drawbacks. If Kelley's formula is expanded to a null hypothesis method (including a correct treatment of the stochastic character of the sample information), the classic method reveals itself as a large sample approximation. We conclude that the classic method is undeservedly regarded inferior by the authors who proposed new indices.

The determination of change in a person subjected to an intervention is an important element of many studies. Usually, the researcher is interested in progress or improvement, for instance as a result of a psychotherapy, although such areas as deterioration of mental functioning, e.g., as a consequence of a medical intervention, may also be a focus of attention. There are various means of characterizing change (see for example, Plewis, 1985). This article does not aim to compare or even discuss the various procedures, but focuses on just one mode of establishing change, i.e., the situation where one attempts to ascertain change, improvement or deterioration, by means of two test assessments, a pretest and a post-test, without the use of a control group.

The difference observed between pretest and post-test is an obvious measure of change. However, only when variables perfectly measure the phenomenon they are supposed to measure is the observed difference really dependable. Unfortunately, the assessment of mental status is always contaminated by effects that preclude a perfect measurement. An observed difference may be partly or even totally due to measurement errors, practice effects, or sample fluctuations. Such influences may also conceal true changes. Naturally, researchers are not interested in changes which can be explained by trivial effects. They are concerned in the first place with the question of whether an observed change is of substantive importance, or has *clinical significance* (Jacobson, Follette, & Revenstorf, 1984) . However, clinical significance is only a matter for discussion if the observed change is *statistically reliable*. How clinical relevance should be established has often been disputed, but at the same time there has been a quest for methods and measures for the determination of statistical reliability of an observed change. These measures, which we, following Jacobson et al. (1984), refer to as *Reliable Change Indices (RC)*, are the topic of this article.

A *RC* is always a ratio. The numerator contains the observed change for a given partici-

Address correspondence to: G.H. Maassen, Utrecht University, Faculty of Social Sciences, Department Methodology and Statistics, P.O.Box 80140, 3508 TC Utrecht, The Netherlands. Tel.: 030-2534765. Fax: 030-2535797. E-mail: g.maassen@fss.uu.nl
Accepted for Publication: January 20, 2000.

pant, corrected for the nuisance effects mentioned above if so desired. A measure which calibrates the disturbance effects has been placed in the denominator. This denominator operates as a criterion: If the numerator exceeds the denominator sufficiently, the nuisance effect can be ruled out with high probability as an alternative explanation for the observed change. Usually, the ratio is regarded as (or transformed into) a standardized normally distributed quantity. If the *RC* exceeds a chosen percentile in the normal distribution, the observed change is taken to be *reliable*.

More recently proposed *RC*'s have become increasingly complicated. This is a consequence of attempts to take the different disturbance effects into account. Initially, only measurement unreliability was dealt with (Christensen & Mendoza, 1986; McNemar, 1962, 1969). In subsequent variants, adjustments were also made for regression to the mean. Hsu (1989), Nunnally and Kotsche (1983), and Speer (1992) all proposed a measure in which the pretest score was replaced by a regression estimate for the true pretest score (from the observed pretest score). This means that information on the group (population or sample) to which the person in question belongs, is also required. These measures can be distinguished according to the delineation of the group, or the standard error used in the denominator.

Hageman and Arrindell (1993), and Zegers and Hafkenscheid (Bruggemans, Van de Vijver. & Huysmans, 1997; Hafkenscheid, 1994; Zegers & Hafkenscheid, 1994) adopt a regression estimate of the true difference from the observed difference. The indices of these authors, too, differ with respect to the criterion in the denominator. Chelune, Naugle, Lüders, Sedlak, and Awad (1993), and McSweeny, Naugle, Chelune, and Lüders (1993) proposed measures in which the practice effect resulting from repetition of measurements has been taken into account. Recently, Bruggemans et al. (1997) proposed a *RC*, which combines this correction with the index of Zegers and Hafkenscheid.

Different *RC*'s may lead to different conclusions concerning the effect of an intervention on a given person, a cause of much confusion. In this article, we demonstrate that this confusion is partly the result of the essentially different principles underlying the *RC*'s. Both principles and their practical consequences are discussed. The drawbacks entailed by the way some of the *RC*'s are constructed will be demonstrated. Several salient *RC*'s are central to our argument, but our reasoning and our remarks apply to other measures as well.

Null Hypothesis Model and Estimation Interval Model

The classic approach

An obvious estimator for the true intra-individual change of a given person i is the observed difference D_i between the pretest score X_i and the post-test score Y_i. In terms of classical test theory (CTT), an observed difference D_i can be split into a true difference and a component containing measurement error:

$$D_i = Y_i - X_i = \Delta_i + E_{D_i}$$

Usually it is assumed that, for all participants, the error components are normally distributed with zero mean and standard deviation equal to σ_{E_D} (Lord & Novick, 1968, p.159). We note that the observed difference score, then, is an unbiased estimator of the true difference. Under the null hypothesis that the treatment has no effect,

$$D_i / \sigma_{E_D} \qquad (1)$$

has a standardized normal distribution. According to McNemar (1962, 1969), an observed difference score is considered *dependable* – the term reliable change had not yet been introduced – if this quantity exceeds a chosen critical value, for instance 1.96. Thus, the procedure can be characterized as a statistical test for the null hypothesis H_0: ''The true change of the given individual equals 0''. There is always a risk that rejection of H_0 is purely an artifact of an unreliable measurement instrument. If alpha is set at .05, the probability of committing a type I error is .05. If a change is called reliable only in the case of improvement ($Y > X$), or in the case of deterioration ($Y < X$), this probability is .025.

The objective of this procedure is to rule out measurement error as an explanation for an observed change with a known but low risk of drawing a wrong conclusion.

It should be noted that this procedure may reflect clinical practice in a simplified way. Firstly, the assumptions of zero mean of error components and equal standard error of measurement of the difference score may be violated in clinical situations. For instance, when a therapeutic client is selected for a low pretest score, the assumption of zero mean of error components is no longer valid. The method is obviously not suitable for such situations. Secondly, this method, as well as the other methods discussed in this article, has to depart from the assumption that effects which may jeopardize the validity of conclusions, such as a practice effect due to repeated measurement, are zero. The procedures that are central to this article cannot be experimentally controlled and there is no way of circumventing this assumption.

This simple procedure, based on statistical testing of a null hypothesis, is central to this article and will be referred to as the *classic approach*. The procedure has been rediscovered by Christensen and Mendoza (1986), and Jacobson and Truax (1991). In both publications, the authors also assume that the standard errors of measurement of pretest and post-test are equal, which is not relevant to our discussion.

The Regressed Score Method

Several authors held that the *RC* could be sharpened by replacing the observed change by an improved estimate for the true change in the numerator. The observed difference is an unbiased estimate of the true change, but other estimates may be worth considering for other reasons. For instance, Nunnally and Kotsch (1983), Hsu (1989), and Speer (1992) emphasized that when estimating the true difference score of person *i*, regression to the mean should be taken into account. Hsu (1989) states, ‘‘Given assumptions of the reliability model, regression toward the mean can be interpreted as the difference between pretreatment and post-treatment scores which would be expected in the absence of treatment effect, because of unreliability of the response measure’’ (p. 462). These authors propose to estimate the post-test score which would be expected in the absence of a treatment effect:

$$\hat{Y}_i = \varrho_{XX}\,(X_i - \mu_X) + \mu_X$$

Here μ_X and ϱ_{XX}, respectively, are the mean and the reliability of the pretest score in a relevant population, called the *reliability group* (Hsu, 1989). Such an estimate for the final score has also been referred to as the *regressed score* (McNemar, 1958, 1969). Then, the improved estimate of the true difference score is taken to be the difference between this estimate for the post-test score and the observed post-test score:

$$\hat{\Delta}_i = Y_i - \hat{Y}_i = D_i + (X_i - \mu_X)\,(1 - \varrho_{XX}). \qquad (2)$$

The standard error of estimate of this estimation is:

$$\sigma_X\sqrt{1-\varrho_{XX}^2} \qquad (3)$$

where σ_X denotes the standard deviation of the pretest score in the reliability group. The *RC* proposed by Hsu (1989) is the ratio with Equation 2 in the numerator and Equation 3 in the denominator, and, in fact, is applied as follows: ‘‘If the estimation interval

$$D_i + (X_i - \mu_X)\,(1 - \varrho_{\mathrm{XX}}) \pm 1.96\ \sigma_X\sqrt{1-\varrho_{XX}^2}$$

does not contain 0, the observed change is called *reliable*.’’ The factor 1.96 suggests that this conclusion has a reliability of 95%.

This procedure suffers from the shortcoming that the estimate according to Equation 2 is generally a biased estimate for the true difference. This is revealed by:

$$\mathrm{E}\,(\hat{\Delta}_i \mid \xi_i, \Delta_i) = \Delta_i + (\xi_i - \mu_\xi)\,(1 - \varrho_{XX}), \qquad (4)$$

where ξ_i denotes the true pretest score of person *i*, and μ_ξ denotes the population mean of the true pretest scores. To express that a given person *i* is not characterized by his or her observed pretest score, which is stochastic, but by the true score, the expectation in Equation (4) is conditioned on the true pretest score as well.

The estimate is only unbiased if $\xi_i = \mu_\xi$ or $\varrho_{XX} = 1$. If Δ_i and $\xi_i - \mu_{\hat{\imath}}$ have a different sign, then Δ_i and the expected value of the estimate can have a different sign. If the bias of the estimate is sufficiently large, it may be expected that a reliable change in the wrong direction will be concluded. This is the case if:

$$|\xi_i - \mu_\xi|\,(1 - \varrho_{XX}) - |\Delta_i| > 1.96\ \sigma_X\sqrt{1-\varrho^2_{XX}}\,.$$

If there is no real change, it may be expected that nevertheless the conclusion ‘reliable change’ will be drawn if:

$$\left|\frac{\xi_i-\mu_\xi}{\sigma_X}\right| > 1.96\ \sqrt{\frac{1+\varrho_{XX}}{1-\varrho_{XX}}}. \tag{5}$$

The classic approach and the method under consideration are based on different principles. The latter can not boast a uniform probability distribution that the researcher can rely on to indicate the probability of making a type I error. In the method in question, an estimation interval is used for the true difference score, based on the regressed score estimate of the true pretest score. Hence, we call this method an *estimation interval method*. At most, the interval that contains the true difference score with high probability is roughly indicated. In the majority of the cases, the decision of reliable change will be justified. However, the estimation of the true difference score is biased, and for patients with an extreme true pretest score (see Equation 5) the bias can be so great that there is a high probability of making a wrong decision. The lower the reliability of the pretest score, the greater is this threat. Hsu (1989) shows in his Table 1, how in such cases the criteria for the decisions in the classic approach and the regressed score method may diverge. If the population mean and variance of the pretest scores are unknown, and if they are estimated with the help of a sample, the regressed score method entails the additional drawback that sample fluctuations may also be conducive to a wrong decision. The regressed score estimate may be attractive as an estimate, but as a basis for the assessment of reliable change it entails shortcomings compared with the null hypothesis model, with no clear advantages.

Reliable Change Indices Based on Kelley’s Formula

Hageman and Arrindell’s RC_{ID}

Several authors have held the view that the *RC* could be improved by sharpening even more the estimate of the true difference in the numerator. Hageman and Arrindell (1993) were the first to employ Kelley’s formula (Kelley, 1947, p.409) as the improved estimate and as the basis of a *RC*. Kelley’s estimate, also known as the *weighted reliability measure*, is:

$$\hat{\Delta}_i = \varrho_{DD}D_i + (1 - \varrho_{DD})\ \overline{D} \tag{6}$$

It is a regression estimate of the true change and thus it has the advantage that, on the whole, the estimation errors are minimized in the sense of least squares (Rogosa, Brandt, & Zimowski, 1982). We note in this context that various regression estimates are eligible. A regressed score estimate was discussed earlier which only employs the pretest score. Kelley’s formula is based on the observed difference. Even more precise are estimates which employ observed initial and final scores separately (see Cronbach & Furby, 1970; Lord, 1956; McNemar, 1958), but as far as we know, these have never been used as a basis for a *RC*.

The question is now: Which denominator fits Kelley’s formula? Hageman and Arrindell (1993) proposed the standard error of measurement of the difference score:

$$\sigma_{\mathrm{E_D}} = \sigma_{E_D} = \sqrt{\sigma^2_{E_X}} + \sigma^2_{E_Y}$$

The *RC* with this standard error in the denominator and Equation 6 in the numerator was named RC_{ID} by Hageman and Arrindell (1993), who simply state that their formula is “the most appropriate way of combining (a) the benefits of the original formulas of Jacobson et al. (1984)

and Christensen and Mendoza (1986), and (b) the search for a better approximation of the true difference score'' (p. 697). Thus, their choice is hardly justified statistically. Nevertheless, Hageman and Arrindell's RC_{ID} seems to have become established in the research literature. It has already been applied by several researchers (Debats, 1996; De Haan et al., 1997; Rudy, Turk, Kubinski, & Zaki, 1995; Van Oppen et al., 1995; Wykes, 1998) or at least mentioned as a serious alternative (Barkham et al., 1996; Taylor, 1995).

The RC of Zegers and Hafkenscheid

As a sequel to Hageman and Arrindell's index, Zegers and Hafkenscheid (1994) proposed the following standard error:

$$\sigma_{\Delta.\mathrm{D}} = \sigma_{\Delta}\sqrt{1-\varrho_{DD}} = \sigma_D\sqrt{\varrho_{DD}}\ \sqrt{1-\varrho_{DD}} = \sigma_{E_D}\ \sqrt{\varrho_{DD}}. \qquad (7)$$

This expression has been known for some time as the *standard error of estimate* belonging to Kelley's formula (Lord & Novick, 1968; McNemar, 1958). Once again, within the context of reliable change indices an old formula has been rediscovered. The index of reliable change with Equation 6 in the numerator and Equation 7 in the denominator was christened RC_{URCI} by Zegers and Hafkenscheid. The authors did not succeed in publishing their *RC* officially and RC_{URCI} found its way into the literature circuitously (see Bruggemans et al., 1997). Recently, Hageman and Arrindell (1999), however, did publish the same index (accompanied by the condition $\varrho_{\mathrm{DD}} > .40$), now referred to as RC_{INDIV}. They seem to recognize their error as they state: ''$\mathrm{RC}_{\mathrm{INDIV}}$ may also be considered an improved version of the $\mathrm{RC}_{\mathrm{ID}}$ index (...) Though under standard conditions, $\mathrm{RC}_{\mathrm{ID}}$ could be considered superior to RC in terms of correct classification of individuals, the present authors now recommend its even more precise successor $\mathrm{RC}_{\mathrm{INDIV}}$,'' (p. 1175). As it seems only fair, throughout this article the new index will be denoted by RC_{URCI}.

Due to its history, RC_{URCI} has hardly been applied yet. Nevertheless, this index deserves more attention, because it certainly has statistical foundation. The method of Zegers and Hafkenscheid is a regression estimate method, where Kelley's formula is (correctly) regarded as an estimate of the true change of person *i*, given his or her observed difference, and where the standard error of estimate in Equation 7 is appropriate. We will now elaborate on the features of this method.

If Zegers and Hafkenscheid choose Equation 7 as the denominator for their *RC*, then in fact they are proposing to use Kelley's formula as a criterion for *reliable change* in the following way:

''The true value of the difference of person *i* is comprised with probability .95 by the interval:

$$\varrho_{DD}D_i + (1-\varrho_{DD})\ \overline{D} \pm 1.96\ \sigma_{\Delta.\mathrm{D}}$$

and if 0 lies outside this interval the observed difference is denominated *reliable*.''

With regard to this approach, we observe the following. Firstly, the procedure has the attractive characteristic that the standard error in Equation 7 is smaller than the standard error of measurement of the difference scores in Equation 1. This can be explained by the fact that when estimating the true difference extra information is used in Equation 6, namely the reliability and the group average of the difference score (see Lord & Novick, 1968, p.68). However, less attractively, the stochastic character of this information has not been taken into account.

Secondly, if the mean true effect of the intervention in the population (μ_{Δ}) is known, then one should obviously take

$$\hat{\Delta}_i' = \varrho_{DD}D_i + (1-\varrho_{DD})\mu_{\Delta}$$

as the estimate of the true change of person *i* rather than Equation 6. This is the estimate with which, on the whole, the errors of estimation are minimized. It is also known that this estimate is generally biased (Willett, 1988), since its mean (expected) value is:

$$\mathrm{E}\,(\hat{\Delta}_i'|\Delta_i) = \varrho_{DD}\Delta_i + (1-\varrho_{DD})\ \mu_{\Delta} = \Delta_i + (1-\varrho_{DD})\ (\mu_{\Delta}-\Delta_i). \qquad (9)$$

The estimate is only unbiased when the true change of person *i* equals the average true change in the population, or when the difference has been perfectly assessed. Usually, neither will be the case. Our comments here are comparable to those regarding the regressed score method. If Δ_i and $\mu_\Delta - \Delta_\mathrm{i}$ have different signs, Δ_i and the expected value of the estimate can have different signs. If the bias of the estimate is sufficiently large, it can be expected that the reliable change is assessed in the wrong direction. This is the case if:

$$|\Delta_\mathrm{i} - \mu_\Delta \mid (1 - \varrho_{DD}) - \mid \Delta_i \mid > 1.96\ \sigma_\mathrm{D} \sqrt{\varrho_{DD}}\ \sqrt{1-\varrho_{DD}}.$$

If no true change of person *i* has taken place, it can nevertheless be expected that the observed change will be taken as reliable if:

$$\left| \frac{\mu_\Delta}{\sigma_D} \right| > 1.96 \sqrt{\frac{\varrho_{DD}}{1-\varrho_{DD}}}. \qquad (10)$$

The lower the reliability of the difference scores, and the more that the true change of person *i* and the average true change in the population differ, the higher is the probability of drawing a wrong conclusion (see Equation 10). The interval of Zegers and Hafkenscheid may be useful as an estimation interval for the true gain of person *i*. However, instead of ruling out trivial explanations for an observed difference, such as the unreliability of the measurements, the procedure opens the possibility that the denomination of an observed difference as reliable may in fact be explicable by other trivial factors.

The classic approach and the approach of Zegers and Hafkenscheid are similar in the sense that both methods are based on the observed change of person *i*. However, the principles underlying the two methods are different and this may even lead to contrary conclusions in practice. In order to demonstrate this, we write RC_URCI as follows:

$$\left| \frac{\varrho_{DD} D_i + (1-\varrho_{DD})\overline{D}}{\sigma_{E_D}\sqrt{\varrho_{DD}}} \right| > 1.96. \qquad (11)$$

We apply the following reparametrization:

$$D_i = A_i * 1.96 \text{ and } \overline{D} = B * 1.96 * \varrho_{E_D}$$

(Thus $A_i > 1$ implies that in the classic approach an observed difference is designated a reliable change.) Now Equation 11 reduces to:

$$\left| A_i * \sqrt{\varrho_{DD}} + \frac{B}{\sqrt{\varrho_{DD}}} * (1-\varrho_{DD}) \right| > 1.$$

Table 1 shows, for given values of *B* and ϱ_{DD}, the minimum values of *A* that lead to the assessment of a reliable change according to RC_URCI. The table shows only the *A* values for positive values *B*. This will do, since the full Table is symmetric: If one changes the sign of *B*, then the sign of *A* also changes. The table covers a realistic value range of *B* (0.5 through 3.0), which can be demonstrated as follows. Incorporating Cohen's *d* reflecting effect size (Cohen, 1977, pp.20, 48) and assuming equal variances S_X and reliabilities ϱ_{XX} of pre- and post-test scores (for the sake of simplicity), the following formula can be derived:

$$B = \frac{\overline{D}}{1.96\sigma_{E_D}} = \frac{dS_X}{1.96\sqrt{2\sigma_E^2}} = \frac{d}{1.96\sqrt{2(1-\varrho_{XX})}}.$$

The largest average effect size of a psychotherapy reported by Smith, Glass, and Miller (1980, p.89) is 2.38 (averaged over 57 effects of cognitive psychotherapy for an unknown number of studies). If Cohen's *d* equals 2.38 and, for instance, $\varrho_{XX} = .92$, then, under the assumptions mentioned, *B* is equal to 3.0.

The section of Table 1 where $A < -1$, draws special attention, for then, in the classic approach, the observed change will be interpreted as a reliable deterioration. (The section where $B < 0$ and $A > 1$, not shown in the table, of course, is of equal importance.) From the table we see, for instance, that when $B = 1.5$ and $\varrho_{DD} = .35$ (which is the case when, say, $\varrho_{XX} = .85$, $\varrho_{XY} = .77$, and $d = 1.61$) following the classic method, the researcher will interpret a negative difference score as a reliable deterioration of person *i*

Table 1. Minimum Values of *A* Leading to the Designation of Reliable Change According to the *RC* of Zegers and Hafkenscheid, For Given Values of *B* and ϱ_{DD}.

ϱ_{DD}	*B*					
	0.5	1.0	1.5	2.0	2.5	3.0
.25	.50	–1.00	–2.50	–4.00	–5.50	–7.00
.30	.66	–.51	–1.67	–2.84	–4.01	–5.17
.35	.76	–.17	–1.10	–2.02	–2.95	–3.88
.40	.83	.08	–.67	–1.42	–2.17	–2.92
.45	.88	.27	–.34	–.95	–1.56	–2.18
.50	.91	.41	–.09	–.59	–1.09	–1.59
.55	.94	.53	.12	–.29	–.70	–1.11
.60	.96	.62	.29	–.04	–.38	–.71
.65	.97	.70	.43	.16	–.11	–.38
.70	.98	.77	.55	.34	.12	–.09
.75	.99	.82	.65	.49	.32	.15

and, at the same time, according to RC_{URCI} as a reliable improvement!

The Reliability-Stability Index Based on RC_{URCI}

Recently, Bruggemans et al. (1997) proposed a *RC* in which a correction for practice effects is combined with the method of Zegers and Hafkenscheid. Basically, the change in every patient of the sample is compared with the mean change in a matched control group. Then, RC_{URCI} is applied to such a control group:

$$RC_{mc}=\frac{D_{mc}\ \varrho_{DD}+D_c(1-\varrho_{DD})}{\sigma_D\sqrt{\varrho_{DD}}\ \sqrt{1-\varrho_{DD}}},$$

where D_{mc} and D_{c} respectively denote the observed change in the matched control group and in the control group as a whole. If RC_{URCI} is also applied with respect to patient *i* (and the result is denoted by RC_{i}), then the reliability-stability index based on RC_{URCI} is defined as follows:

$$RST_{\mathrm{new}} = RC_{\mathrm{i}} - RC_{\mathrm{mc}}.$$

Bruggemans et al. argue that "because RST_{new} is the difference between two *z*-scores, like for the other *z*-score indices, its absolute value had to be larger than 1.645 in order to indicate significant cognitive deterioration" (p. 548). (The authors were interested in the negative effects of a medical intervention.) This procedure compounds several mistakes. First, it has been shown above that neither component should be regarded as a *z*-score if the change in patient *i* differs from the average change in the population. Secondly, the difference of two *z*-scores generally is not a *z*-score. This is only the case when the correlation between the two components equals .5. If this correlation is higher, then the standard deviation of the difference is smaller than 1. Thirdly, the stochastic character of the results in the control group(s) is not taken into account. It should not be surprising that the conclusions drawn with RST_{new} differ strongly from the conclusions resulting from other criteria. Bruggemans et al. report relatively many reliable changes on the basis of their RST_{new}. A trivial explanation for this finding may be the correlation between both components. It will be clear that the use of RST_{new} is not to be recommended.

Kelley's Formula as the Basis for a Null Hypothesis Method

We have shown that the null hypothesis method is clear in the sense that the appropriate null hypothesis "$\Delta_{\mathrm{i}} = 0$" is put to a statistical test. In this way, measurement unreliability can be ruled out with high probability as an explanation for the observed difference score. It was emphasized that Kelley's formula has its advantages as an estimate, but the way it has been expanded into a *RC* has serious drawbacks.

These observations raise the question of whether Kelley's formula can serve as the basis of a null hypothesis method with the concomitant advantages. In answering this question, we distinguish two situations. First, we start from the practical situation that the average true change in the population μ_Δ is unknown. In the second situation, it will be assumed that the researcher knows this quantity.

Average true change in the population unknown
In the form of Equation 6, also, Kelley's formula is a biased estimate of the true change of person *i* (Maassen, 1998, 2000; Rogosa et al., 1982). Again, the mean (expected) value is given by Equation 9. If the fact that the sample information is stochastic is taken into account, the standard error of estimate of the predicted true score is not the correct denominator for the *RC*. The variance of Kelley's estimate should be adopted, with the sample information treated as a random element. We have calculated this variance elsewhere (Maassen, 1998, 2000):

$$Var\,(\hat{\Delta}_{\mathrm{i}}|\Delta_{\mathrm{i}}) = \sigma^2_{E_D}\left(\varrho^2_{DD}+\frac{(1+2\varrho_{DD})(1-\varrho_{DD})}{n}\right). \qquad (12)$$

(Here it is assumed that person *i* belongs to the sample whose information is used in Kelley's formula.) In order to yield a standardized normally distributed variable, Equation 9 is subtracted from Equation 6 and the result is divided by the square root of Equation 12. If, finally, the null hypothesis "$\Delta_{\mathrm{i}} = 0$" is implemented, the following expression results:

$$\frac{\varrho_{DD}D_i+(1-\varrho_{DD})\;(\overline{D}-\mu_\Delta)}{\sigma_{E_D}\sqrt{\varrho^2_{DD}+\frac{(1+2\varrho_{DD})\;(1-\varrho_{DD})}{n}}}. \qquad (13)$$

Under the null hypothesis, this expansion of Kelley's estimate has a standardized normal distribution and can be regarded as a *RC* in the shape of a test statistic. Equation 13, however, is of theoretical interest rather than practical importance. On the one hand, it offers the same advantages as the classic approach, such as the testing of an appropriate null hypothesis and a limited chance of committing a Type I error. Moreover, the bias of Kelley's estimate and the stochastic character of the sample information are adequately treated (which inevitably entails a certain loss of power). On the other hand, application of this *RC* requires information on the effect of the intervention in the population (μ_Δ), which above was assumed to be unknown. The researcher then has to estimate μ_Δ by means of a large sample. Let us assume that the sample size is sufficiently large for the sample statistics to be a good approximation of the population values. Then, from Equation 12 follows:

If n → ∞, then $\overline{D} - \mu_\Delta \rightarrow 0$ and $\mathrm{Var}(\hat{\Delta}_{\mathrm{i}}|\Delta_{\mathrm{i}}) \rightarrow \varrho^2_{DD}\;\sigma^2_{E_D}$.

This implies a return to the start: The classic method. Consequently, the classic approach can be regarded as the large sample approximation for the properly composed *RC* based on Kelley's formula.

Average true change in the population known
When the average true change in the population is already known, Kelley's formula takes the shape of Equation 8. The expected value of this statistic can also be found in Equation 9, and since the second term on the right of Equation 8 is a constant, its conditional variance is:

$$\mathrm{Var}(\hat{\Delta}'_i|\Delta_i) = \varrho^2_{DD}\;\sigma^2_{E_D}.$$

Standardization of the estimator yields the following standardized normally distributed statistic:

$$\frac{\varrho_{DD}\;(D_i-\Delta_i)}{\varrho_{DD}\sigma_{E_D}} = \frac{D_i-\Delta_i}{\sigma_{E_D}},$$

which under "H_0: $\Delta_{\mathrm{i}} = 0$" boils down to Equation 1, the classic approach. Again, we return to our starting point, which should not surprise the reader.

SUMMARY AND DISCUSSION

Scrutinizing the reliable change indices proposed in the clinical psychology literature over the past twenty years does not encourage confidence in the progress of science. One observes a series of proposals that merely repeat pe-existing knowledge, faulty proposals that have had to be corrected, or erroneous proposals that have not yet been corrected. This has happened in spite of the peer review system. The present author was amazed to find that even when reviewers agree that a previously proposed index is wrong, they may give low priority to an adjustment. Paraphrasing Churchill, it would seem that ''the peer review system is the worst, save all other systems.''

Explanations for the lack of continuous progress in the development of methods for the assessment of reliable change are not difficult to find. A lack of statistical knowledge and an uncritical zeal to demonstrate the efficacy of an intervention, which should require constant refining of the criteria employed, are undoubtedly conducive to the current situation.

In this article, a number of previously proposed reliable change indices (*RC*'s) are examined. According to the principle which underlies them, they can be categorized as a null hypothesis testing or an estimation interval method. The only representative of the former category discussed here is based on the ratio of the observed difference and the standard error of measurement of the difference scores. With this method, which we have named the classic approach, the probability that observed changes erroneously will be designated reliable is limited by a low level of significance.

Most of our attention has been devoted to various estimation interval methods. These methods originate in the view that the estimate of the true change should be improved. We have elaborated on the two representatives of this category that possess an appropriate standard error in the denominator, namely *the regressed score method*, proposed among others by Hsu, and the *RC of Zegers and Hafkenscheid*. In the majority of the cases, all the methods mentioned will probably lead to the same interpretation of the observed change, but in this article it has been shown that, with regard to the estimation interval methods, no uniform upper limit exists for the probability of an incorrect conclusion. This probability depends on the relative position of the patient within the population to which he or she belongs, either with respect to the true initial score (in the regression estimate method) or the true difference score (in the method of Zegers and Hafkenscheid). The probability of committing a type I error may be high if the participant in question takes an extreme position in the population. Only a few participants will be the victim of a type I error or an interpretation in the wrong direction, but for deviant individuals a faulty interpretation may have particularly important consequences. Whereas the classic method rules out with high probability measurement errors as a possible explanation for an observed change, the qualities of the estimation interval methods are less clear. Trivial effects may possibly lead to an observed change (or an observed zero change) being designated reliable, or an observed change interpreted as reliable change in the wrong direction. Such trivial effects also include sample fluctuations, if the *RC* comprises sample information.

Recently, various estimation interval-based *RC*'s have been proposed, which depart from Kelley's estimate of the true change of a given person. The RC_{URCI} of Zegers and Hafkenscheid is the most salient example. We observe that some individuals run the risk that the classic approach and RC_{URCI} can lead to the designation of reliable change in opposite directions. Considering our earlier observations, we put our trust in the results of the classic approach rather than the estimation interval method.

Finally, we have examined whether Kelley's estimate may be the basis of a null hypothesis method. This led us to the following conclusions: (1) If Kelley's formula is correctly expanded to a *RC*, the researcher is confronted with the problem that, in principle, population information is required; this information is generally not available; (2) If the researcher has a large sample from the population at his or her disposal, the sample or population information turns out to be superfluous. The classic approach reveals itself as the large sample approximation for correctly composed *RC*'s based on Kelley's formula.

All in all, in our view there are strong arguments for preferring the classic approach to the estimate interval method. This method has been undeservedly regarded inferior by the authors who recently proposed new indices in the clinical psychology literature.

REFERENCES

Barkham, M., Rees, A., Stiles, W.B., Shapiro, D.A., Hardy, G.E., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64*, 927-935.

Bruggemans, E., Van de Vijver, F.J.R., & Huysmans, H.A. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology, 19*, 543-559.

Chelune, G.J., Naugle, R.I., Lüders, H., Sedlak, J., & Awad, I.A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*, 41-52.

Christensen, L., & Mendoza, J.L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 12*, 305-308.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cronbach, L.J. & Furby, L. (1970). How we should measure "change" – or should we? *Psychological Bulletin, 74*, 68-80.

Debats, D.L. (1996). Meaning in life: Clinical relevance and predictive power. *British Journal of Clinical Psychology, 35*, 503-516.

De Haan, E., Van Oppen, P., Van Balkom, A.J.L.M., Spinhoven, P., Hoogduin, K.A.L., & Van Dyck, R. (1997). Prediction of outcome and early vs. late improvement in OCD patients treated with cognitive-behavior therapy and pharmacotherapy. *Acta Psychiatrica Scandinavica, 96*, 354-361.

Hafkenscheid, A.J.P.M. (1994). *Rating scales in treatment efficacy studies: Individualized and normative use*. Unpublished doctoral dissertation, Rijksuniversiteit Groningen. Groningen, The Netherlands.

Hageman, W.J.J.M., & Arrindell, W.A. (1993). A further refinement of the reliable change (RC) index by improving the pre-post difference score: Introducing RC_{ID}. *Behaviour Research and Therapy, 31*, 693-700.

Hageman, W.J.J.M., & Arrindell, W.A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy, 37*, 1169-1193.

Hsu, L.M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment, 11*, 459-467.

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 Clinical and Consulting Psychology, 59*, 12-19.

Kelley, T.L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press.

Lord, F.M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421-437.

Lord, F.M., & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Maassen, G.H. (1998). The reliability weighted measure of individual change as an indicator of reliable change. *Kwantitatieve Methoden, 19*, nr. 58, 29-40.

Maassen, G.H. (2000). Kelley's formula as a basis for the asessment of reliable change. *Psychometrika, 65*, 187-197.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement, 18*, 47-55.

McNemar, Q. (1962, 3rd ed.). *Psychological Statistics*. New York: Wiley.

McNemar, Q. (1969, 4th ed.). *Psychological Statistics*. New York: Wiley.

McSweeny, A.J., Naugle, R.I., Chelune, G.J., & Lüders, H. (1993). "T Scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist, 7*, 300-312.

Nunnally, J.C., & Kotsch, W.E. (1983). Studies of individual subjects: Logic and methods of analysis. *British Journal of Clinical Psychology, 22*, 83-93.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726-748.

Rudy, T.E., Turk, D.C., Kubinski, J.A., & Zaki, H.S. (1995). Differential treatment responses of TMD patients as a function of psychological characteristics. *Pain, 61*, 103-112.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). *The Benefits of Psychotherapy*. Baltimore: Johns Hopkins University Press.

Speer, D.C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60*, 402-408.

Taylor, S. (1995). Assessment of obsessions and compulsions: Reliability, validity and sensitivity to treatment effects. *Clinical Psychology Review, 15*, 261-296.

Van Oppen, P., de Haan, E., Van Balkom, A.J.L.M., Spinhoven, P., Hoogduin, K., & Van Dyck, R. (1995). Cognitive therapy and exposure in-vivo in the treatment of obsessive-compulsive disorder. *Behaviour Research and Therapy, 33*, 379-390.

Willett, J.B. (1988). Questions and answers in the measurement of change. In E.Z. Rothkopf (Ed.): *Review of research in education, 15 (1988-89)*, 345-422. Washington: American Educational Research Association.

Wykes, T. (1998). What are we changing with neurocognitive rehabilitation. Illustrations from 2 single cases of changes in neuropsychological performance and brain systems as measured by SPECT. *Schizophrenia Research, 34*, 77-86.

Zegers, F.E., & Hafkenscheid, A.J.P.M. (1994). The ultimate reliable change index: An alternative to the Hageman & Arrindell approach. Groningen: *Universiteit van Groningen, Heymans Bulletin HB-94-1154-EX*.