The Reliability-Weighted Measure of Individual Change as an Indicator of Reliable Change*

by

Gerard H. Maassen

Abstract

In the literature on the measurement of change, *reliable change* is usually determined by means of a confidence interval around an observed value of a statistic that estimates the *true change*. The definition of such an interval is normally based on the assumption that the statistic has a standardized normal distribution. In the recent literature, attention has been particularly directed to the improvement of the estimation of the true change. However, when authors fail to incorporate an adequate standard error of the estimator, the assumption of a standardized normal distribution is no longer justified. Consequently, the statistical characteristics of the resulting confidence interval are unclear. In this article these points are demonstrated with respect to the Reliable Change Index, incorporating the *reliability-weighted measure of individual change*, recently proposed by Hageman and Arrindell (1993).

*The author would like to thank Coen Bernaards and Klaas Sijtsma for their comments on an earlier draft of this paper.

---

dr. G.H. Maassen
Utrecht University
Faculty of Social Sciences
Department Methodology and Statistics
Post Box 80140
3508 TC Utrecht, The Netherlands
tel: 030-2534765, fax: 030-2535797
e-mail: g.maassen@fsw.ruu.nl

1. Introduction

In almost any social-scientific field of research, the assessment of change - or more specifically gain or growth - is an important topic. In many situations a researcher may wish to examine whether an observed difference between two measurements for a subject on a certain variable is dependable, i.e. greater than expected on the basis of errors of measurement.

For some time a simple procedure for asserting the dependability of an observed change has been known (see McNemar, 1962, p.154, and the next section). In psychotherapy research, where the assertion of the effect of treatments is of specific importance, several attempts have been made to sharpen the procedure. In that field, criteria for the assessment of dependable change are usually called indices of *reliable change* (RC) and the psychotherapy research literature shows an interesting development in the adaptation of these indices. Jacobson, Follette and Revenstorf (1984) proposed an index based on the observed difference between a pretest and a posttest measurement of a given patient, but once Christensen and Mendoza (1986) had pointed it out, the earlier authors recognized that the standard error in their formula was incorrect. Christensen and Mendoza (ibid.) re-introduced a procedure that is also based on the observed difference without noticing that this boiled down to the classic approach to which we referred earlier. Nunnally and Kotsch (1983), Hsu (1989), Speer (1992) and Hageman and Arrindell (1993) all advocated variants in which the observed difference is replaced by an estimation of the true change that they considered more reliable. Hageman and Arrindell incorporated the *reliability-weighted measure of individual change* (Willett, 1988) in their formula (see next section).

In this article, first the composition of the most recently proposed index of reliable change, that of Hageman and Arrindell, is examined. Finally, the question is discussed of what advantage adapting the reliability-weighted measure of individual change has to offer above the classic approach. In our view, the relevance of this question exceeds the boundaries of psychotherapy research.

2. The Observed Change and the Reliability-Weighted Measure of Individual Change

According to McNemar (1962, 1969), an observed difference score is considered dependable if:

$$|D_i| = |Y_i - X_i| > 1,96\ \sigma_{e_d}, \qquad (1)$$

where X_i and Y_i respectively are the initial and final score of a given person i on a certain test and σ_{e_d} the standard error of (measurement of) the difference score. (The way this standard error is calculated in an actual situation is not relevant to our argument.) In this article this method plays a central role and will be referred to as the *classic approach*. The approach is based on the following argument. In terms of the classical test theory (CTT) an observed difference D_i can be split into a true difference and a component containing measurement error:

$$D_i = Y_i - X_i = \Delta_i + e_{d_i}. \qquad (2)$$

The application of difference scores as an estimator for the true difference is usually simplified by assuming that for all subjects the error components are normally distributed with zero mean and standard deviation equal to σ_{e_d} (see Lord & Novick, 1968, p.159). For a given (not randomly selected) person i, i.e. for a fixed value of Δ_i, the following holds:

$$\mathcal{E}(e_{d_i}|\Delta_i) = 0 \quad thus \quad \mathcal{E}(D_i|\Delta_i) = \Delta_i$$

$$Var(e_{d_i}|\Delta_i) = \sigma^2_{e_d} \quad thus \quad Var(D_i|\Delta_i) = \sigma^2_{e_d}.$$

Under the null hypothesis that the treatment has no effect, D_i / σ_{e_d} has a standardized - normal distribution. If a change is designated reliable in the sense of form.(1), there is always a risk that it is purely an artifact of an unreliable measurement instrument. The probability of committing a type I error is .05. If a change (improvement) is called reliable only in case $Y > X$, this probability is .025.

We note that the observed difference score is an unbiased estimator of the true difference. In recent literature, authors have advocated the implementation of estimators that are considered preferable to the observed difference. For instance, Hageman and Arrindell (1993) proposed their *Reliable Change Index, utilizing improved difference scores* (RC_{ID}), in which information of a sample is incorporated. Applying by the symbols used by Hageman & Arrindell, this index is defined as follows:

$$RC_{ID} = \frac{(x_2 - x_1) r_{DD} + (M_2 - M_1)(1 - r_{DD})}{\sqrt{S^2_{E(1)} + S^2_{E(2)}}}, \qquad (3)$$

where x_1 and x_2 respectively are the pretest and posttest score of a given subject; M_1 and M_2 are respectively the pretest and posttest means of a sample of subjects who have received the treatment; r_{DD} is the reliability of the difference scores. A change is designated reliable by Hageman and Arrindell if the absolute value of RC_{ID} exceeds 1.96. Using symbols introduced above the definition becomes:

$$|\hat{\Delta}_i| = |\rho_{dd} D_i + (1 - \rho_{dd}) \overline{D}| > 1{,}96\sigma_{e_d}. \qquad (4)$$

With regard to the conversion into our symbols we note that the reliability is a population parameter, not a sample statistic; we prefer to use a Greek symbol. In the denominator of form.(3) the standard error of the difference is broken down into the standard errors of measurement of the separate observations. This itemization is not expedient for our argument and is therefore not carried through to form.(4).

The numerator of Hageman and Arrindell's index contains what is called by Willett (1988) the *reliability-weighted measure of individual change*. It is readily apparent that the numerator is a weighted mean of an individual observed difference and the mean difference observed in the sample. In fact, it is the *regression function of the true difference on observed difference*, in which the population mean of differences is replaced by a large-sample mean (Lord & Novick, 1968, p.152). The regression function of the true difference on observed difference minimizes the expected squared error of estimation (Lord & Novick, 1968, p.65; Rogosa, Brandt & Zimowski, 1982). Thus, in this respect, the reliability-weighted measure of individual change may be considered to be an improved estimator (compared with the observed difference) of the true improvement of person *i*. It is, however, not an unbiased estimator (Rogosa et al., 1982, p.736; Willett, 1988, p.379), i.e. $\mathcal{E}(\hat{\Delta}_i|\Delta_i) \neq \Delta_i$. This characteristic, which is of central importance in this context, will be apparent from form.(6) (see below).

Hageman and Arrindell give no justification for the denominator of their index in terms of a standard error of measurement of the statistic in the numerator. The denominator of form.(3) does not contain the standard error of the numerator but that of D_i. This fact and the fact that the numerator is a biased estimator of the true change together make it impossible to make any statement about the probability of making a type I error, which is a requirement

of any confidence interval. Their statement that a change is deemed reliable when the absolute value of RC_{ID} exceeds 1.96, suggests that they have a standardized normal distribution for RC_{ID} in mind and a limit of .05 for the risk of committing a type I error (with two-tailed testing of the null hypothesis), but this is not statistically justified.

Below, we will bring the numerator and denominator of form.(3) into correspondence, converting the reliability-weighted measure of individual change into a statistic with a standardized normal distribution, and then review the usefulness of this statistic as an index of reliable change.

We start from form.(4) and split the observed statistics into true and measurement error components:

$$\hat{\Delta}_i = \rho_{dd}\Delta_i + \rho_{dd}e_{d_i} + (1 - \rho_{dd})(\overline{\Delta} + \overline{e}_d) . \qquad (5)$$

In this expression, $\overline{\Delta}$ and $\overline{e}_d$ are the sample means of the true differences and of the measurement errors of the differences, respectively. Thus, $\overline{\Delta}$ (supposedly calculated from a random sample) and $\hat{\Delta}_i$, e_{d_i}, $\overline{e}_d$ are random variables. For our further argument we do not need to assume that $\overline{e}_d = 0$, as is habitual in CTT. We assume that $\rho(e_{d_i}, e_{d_j}) = 0$ for $i \neq j$. Hageman and Arrindell do not make explicit whether they consider the given subject i as a member of the sample or not. We admit the possibility that he or she does indeed belong to the sample, which implies that $\rho(e_{d_i}, \overline{e}_d) \neq 0$.

For the derivation of the parameters of the conditional probability distribution of $\hat{\Delta}_i$ (under the condition that the true difference score of respondent i is Δ_i) we use well-known statistical formulae for the moments of the distribution of a random variable v, having a joint sampling distribution with random variable w (Lord & Novick, 1968, p.35; Rao, 1973, p.97). These formulae are:

$$\mathcal{E}\, v = \mathcal{E}_w\, \mathcal{E}(v|w) \quad \text{and}$$

$$Var(v) = \mathcal{E}_w[Var(v|w)] + Var_w[\mathcal{E}(v|w)] .$$

If the mean of the true difference scores in the sampled population equals μ_Δ then we find for the expected value:

$$\mathcal{E}(\hat{\Delta}_i|\Delta_i) = \mathcal{E}_{\bar{\Delta}}\mathcal{E}(\hat{\Delta}_i|\Delta_i, \bar{\Delta}) = \mathcal{E}_{\bar{\Delta}}[\rho_{dd}\Delta_i + (1 - \rho_{dd})\bar{\Delta}] = \rho_{dd}\Delta_i + (1 - \rho_{dd})\mu_{\Delta}. \quad (6)$$

Starting from

$$Var(\hat{\Delta}_i|\Delta_i) = \mathcal{E}_{\bar{\Delta}}[Var(\hat{\Delta}_i|\Delta_i, \bar{\Delta})] + Var_{\bar{\Delta}}[\mathcal{E}(\hat{\Delta}_i|\Delta_i, \bar{\Delta})], \quad (7)$$

we find for the variance (see Appendix):

$$Var(\hat{\Delta}_i|\Delta_i) = \sigma^2_{e_d}\left[\rho^2_{dd} + \frac{(1 + 2\rho_{dd})(1 - \rho_{dd})}{n}\right]. \quad (8)$$

We note that the quantity

$$\frac{\hat{\Delta}_i - \mathcal{E}(\hat{\Delta}_i|\Delta_i)}{\sqrt{Var(\hat{\Delta}_i|\Delta_i)}}$$ approximately follows a standardized normal distribution.

If the results of form.(6) and form.(8) are substituted in this expression, it is readily apparent that, under the null hypothesis that the true change of person *i* equals zero, the same holds for:

$$\frac{\rho_{dd}D_i + (1 - \rho_{dd})(\bar{D} - \mu_{\Delta})}{\sigma_{e_d}\sqrt{\rho^2_{dd} + \frac{(1 + 2\rho_{dd})(1 - \rho_{dd})}{n}}}. \quad (9)$$

An assessed change may be called reliable if the absolute value of the numerator of form.(9) exceeds 1,96 times the denominator. Now we can confidently assert that the probability of committing a type I error is less than .05.

Form.(9) demonstrates that Hageman and Arrindell's RC_{ID} fails to be standardized normally distributed: Since the reliability-weighted measure of individual change is not an unbiased estimate of the true change, the numerator has to be modified; the denominator of form.(9) contains the appropriate standard error of measurement. Form.(9), however, has more theoretical than practical value. Its greatest disadvantage is that the population mean of the true gain scores must be known, which will almost never be the case. Generally, the researcher will be inclined to estimate this mean from the average of the differences observed in a sufficiently large sample. However, if $n \rightarrow \infty$, we see that:

$\bar{D} - \mu_{\Delta} \approx 0$ and $Var(\hat{\Delta}_i|\Delta_i) \approx \rho^2_{dd}\,\sigma^2_{e_d}$.

Thus, the index of the classic approach (form.(1)) proves to be a large sample approximation for the index that is based on the reliability-weighted measure of individual change! This finding and the obvious advantage that no sample information is required, revalues the classic approach (again).

3. The classic approach and RC_{ID} compared

An inspection of form.(4) shows that the numerator is normally too high in comparison with form.(9). The same holds for the standard error of measurement of the differences used in form.(4). If both effects counterbalance one another, then the classic approach and RC_{ID} will yield approximately the same results. We need to see how far this is the case. In order to gain more insight into the practical implications of both options, we write RC_{ID} somewhat differently:

$$\frac{D_i + (1 - \rho_{dd})(\bar{D} - D_i)}{\sigma_{e_d}} > 1.96. \qquad (10)$$

If an observed difference of a given person equals the mean difference observed in the sample, both procedures lead to identical conclusions. Form.(10) shows that in all other cases application of RC_{ID} implies a 'correction' for the difference from the sample mean. Hageman and Arrindell's index is less conservative than the classic approach if: $\bar{D} - D_i > 0$.
Subjects with an observed difference score exceeding the sample mean are 'unlucky': their difference score is adjusted in a negative direction. That is, Hageman and Arrindell's procedure is then more conservative than the classic approach. The balance of the adjustments depends on the distribution of the difference scores in the sample. If, for instance, the distribution is positively skewed (many extremely high positive changes), then the sample mean exceeds the sample median, which means that more than 50% of the differences are less than the sample mean. In that case, more differences will be 'corrected' upward than downward. If the sample distribution is negatively skewed, the balance is tipped in the opposite direction. Whether Hageman and Arrindell's method is less or more conservative than the classic approach depends on the *shape* of the distribution of the difference scores in the sample.

The fact that RC_{JD} is based on the reliability-weighted measure of individual change implies that the impact of the sample mean of the difference scores is dependent on ρ_{dd}. If ρ_{dd} is very low, the global sample information will overshadow the individual information of the particular subject. In the extreme case that ρ_{dd} equals zero, the individual change score no longer matters. This may be acceptable practice when estimating the subject's true change, but a researcher who uses RC_{JD} must be aware that this characteristic of the reliability-weighted measure of individual change used in combination with the statistically unjustified denominator of RC_{JD} may lead to strange conclusions. In order to facilitate a demonstration of the possible practical implications, we write form.(4) once more in a different way. First we express the observed differences as multiples of the standard error:

$$D = a*1,96\sigma_{e_d} \quad and \quad \overline{D} = b*1,96\sigma_{e_d}$$

(Thus $a > 1$ implies that according to the classic approach an observed difference is designated a reliable change.) Now form.(4) reduces to:

$$|a*\rho_{dd} + b*(1 - \rho_{dd})| > 1.$$

Table 1 shows, for given values of b and ρ_{dd}, the minimum value of a that lead to the assessment of a reliable change according to Hageman and Arrindell's approach. The Table covers realistic value ranges of the reliability of the difference score (.50 through .75) and of b (0.5 through 3.0). With regard to b, incorporating Cohen's d reflecting effect size and assuming equal variances and reliabilities of pre- and posttest scores (for the sake of simplicity), a value range can be derived from the following formula:

$$b = \frac{\overline{D}}{1.96\ \sigma_{e_d}} = \frac{d\ s_x}{1.96\sqrt{2\sigma_e^2}} = \frac{d}{1.96\sqrt{2(1 - \rho_{xx})}},$$

If ρ_{xx} ranges from .70 through .90 and Cohen's d ranges from 1.5 through 2.5, it can be verified from substitution that b then ranges from 0.99 through 2.85.

Apparently, when the reliability of the difference scores is low and/or the mean sample difference is high (b high), RC_{JD} can change an observed negative difference (a negative) into a 'reliable improvement'. A given observed difference can even be interpreted as a reliable deterioration according to the classic approach and at the same time as a reliable improvement according to Hageman and Arrindell's approach! (This is, for instance, the case when the mean difference in the sample equals $3*1.96*\sigma_{e_d}$ and $\rho_{dd} \leq 0.5$.) This is particularly

paradoxical for the subgroup of individuals in the sample who had a low pretest and a negative difference score, which indicates a substantive deterioration because, on the basis

Table 1.

Minimum values of *a* leading to the designation of reliable change according to RC_{ID}, for given values of *b* and ρ_{dd}.

ρ_{dd}	b					
	0.5	1.0	1.5	2.0	2.5	3.0
.50	1.50	1	0.50	0.00	-0.50	-1.00
.55	1.41	1	0.59	0.18	-0.23	-0.64
.60	1.33	1	0.67	0.33	0.00	-0.33
.65	1.27	1	0.73	0.46	0.19	-0.01
.70	1.21	1	0.79	0.57	0.36	0.14
.75	1.17	1	0.83	0.67	0.50	0.33

of regression to the mean, an improvement in observed scores was expected. When a subject's observed change significantly exceeds the magnitude expected on the basis of measurement error, it is simply not conceivable that this should be interpreted as a reliable change in the opposite sense if all that has happened is that he or she has not not benefited from the same treatment as other subjects.

4. Summary

We have shown that, from a statistical point of view, Hageman and Arrindell's RC_{ID} is not justified. For this index, the probability of designating an observed change as reliable, when in fact it is an artifact of an unreliable measurement instrument, is unknown. We have also shown that this index can contain paradoxical implications.

We do not intend to dispute the favorable characteristics of the *reliability-weighted measure of individual change* as an estimator of a subject's true change Δ_i. However, the use of this measure for the assessment of reliable change, in the way that Hageman and Arrindell intended, requires conversion into a statistic whose probability distribution is known. We have shown that an appropriate adjustment of RC_{ID} into a normally distributed statistic (cf.

form.(9)) leads to a procedure which, for increasing n, converges to the classic approach form.(1) for ruling out the unreliability of the measurement instrument as a plausible competing explanation. The classic approach, which in the literature on therapeutic research is also known as the index of Christensen and Mendoza (1986), has an obvious advantage: no sample information is required. This approach is preferable to Hageman and Arrindell's RC_{ID}, which should not be used.

Appendix

Derivation of the standard error of the reliability-weighted measure of individual change.

We start from

$$Var(\hat{\Delta}_i|\Delta_i) = \mathscr{E}_{\bar{\Delta}}[Var(\hat{\Delta}_i|\Delta_i, \bar{\Delta})] + Var_{\bar{\Delta}}[\mathscr{E}(\hat{\Delta}_i|\Delta_i, \bar{\Delta})]. \qquad (7)$$

For the first term on the right side of this expression can be written:

$$\mathscr{E}_{\bar{\Delta}}[Var(\hat{\Delta}_i|\Delta_i, \bar{\Delta})] =$$

$$\mathscr{E}_{\bar{\Delta}}[\rho_{dd}^2 \sigma_{e_d}^2 + (1 - \rho_{dd})^2 \sigma_{\bar{e}_d}^2 + 2\rho_{dd}(1 - \rho_{dd}) cov(e_{d_i}, \bar{e}_d)] =$$

$$\rho_{dd}^2 \sigma_{e_d}^2 + \frac{(1 - \rho_{dd})^2}{n} \sigma_{e_d}^2 + \frac{2\rho_{dd}(1 - \rho_{dd})}{n} \sigma_{e_d}^2,$$

and for the second term on the right side of form.(7):

$$Var_{\bar{\Delta}}[\mathscr{E}(\hat{\Delta}_i|\Delta_i, \bar{\Delta})] = Var_{\bar{\Delta}}[\rho_{dd}\Delta_i + (1 - \rho_{dd})\bar{\Delta}] = \frac{(1 - \rho_{dd})^2}{n} \sigma_{\Delta}^2.$$

Substituting the equality:

$$\frac{(1 - \rho_{dd})^2}{n} (\sigma_{e_d}^2 + \sigma_{\Delta}^2) = \frac{(1 - \rho_{dd})}{n} \sigma_{e_c}^2$$

reduces form.(7) to:

$$Var(\hat{\Delta}_i|\Delta_i) = \sigma^2_{\theta_d} \left[\rho^2_{dd} + \frac{(1 - \rho_{dd})}{n} + \frac{2\rho_{dd}(1 - \rho_{dd})}{n}\right]. \quad (8)$$

(The last term on the right side is cancelled if person i and the sample are independently selected.)

References

Christensen, L., and Mendoza, J.L. (1986). A Method of Assessing Change in a Single Subject: An Alteration of the RC Index. *Behavior Therapy, 17*, 305-308.

Hageman, W.J.J.M., and Arrindell, W.A. (1993). A Further Refinement of the Reliable Change (RC) Index by *I*mproving the Pre-Post *D*ifference Score: Introducing RC_{ID}. *Behaviour Research and Therapy*, 31, 693-700.

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., and Revenstorf, D. (1984). Psychotherapy Outcome Research: Methods for Reporting Variability and Evaluating Clinical Significance. *Behavior Therapy, 15*, 336-352.

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

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

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

Nunnally, J.C., and Kotsch, W.E. (1983). Studies of Individual Subjects: Logic and Methods of Analysis. *British Journal of Clinical Psychology, 22*, 83-93.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley

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

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

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.

ontvangen: 10-06-1997
geaccepteerd: 18-03-1998