PERGAMON

Behaviour Research and Therapy 39 (2001) 495–498

BEHAVIOUR RESEARCH AND THERAPY

www.elsevier.com/locate/brat

The unreliable change of reliable change indices

Gerard H. Maassen

Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, PO Box 80140, 3508 TC, Utrecht, The Netherlands

Received 22 December 1999

Abstract

The classic method for assessment of reliable change, in 1991 re-introduced as Jacobson's RC, can be characterized as a confidence interval method. In recent years, several RC indices have been proposed using Kelley's (1947) (Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press) formula for estimating true change. In these proposals, interval estimation and confidence intervals are mixed up, which leads to unjustified probability statements. When Kelley's estimate is correctly expanded into a normal distributed statistic, the classic approach reveals itself as a large sample approximation of a properly constructed RCI based on Kelley's formula. Researchers should continue using the classic approach for the determination of reliable change.

Keywords: Change scores; Reliable change indices; Kelley's estimate

In his congenial commentary on Hageman and Arrindell (1999), in which new measures for reliable change are once more proposed, David Speer (1999) writes: "There is *no* agreement or consensus among methodologists about the ubiquitousness regression to the mean, its effects on *d*-scores, whether or not *d*-scores are biased and/or reliable and whether or not *d*-scores really need adjustment or correction in the analysis of two-wave data". He argues in favour of "(a) a 10-year moratorium on the use of *adjusted* RC methods in the area of outcome evaluation and (b) that we agree to use RC and only RC during this period".

In this reply I wish to endorse his plea. I will demonstrate (a) that the situation is even worse than perhaps he imagined, and (b) that there is more solid statistical support for his argument than he may have supposed.

Let me first provide an account of the history as I perceive it. As early as 1962, McNemar (1962) proposed a method to establish whether a difference score can be considered *dependable* — the term reliable change had not yet been introduced. The central criterion in this method was the ratio of the observed change and the standard error of measurement of the difference. I am not sure that he was the first to propose this method, nor that this was the first publication of his

0005-7967/01/$ - see front matter
PII: S0005-7967(00)00029-2

proposal. As a precaution, I call this method the *classic approach*. Jacobson, Follette and Ravenstorf (1984) introduced the term *Reliable Change Index* (RCI) and proposed an index with a denominator slightly different from the classic approach. Christensen and Mendoza (1986) showed that their formula was wrong and proposed a formula in a rather ill-chosen notation, which boiled down to the classic approach. Jacobson and his colleagues (Jacobson & Truax, 1991) acknowledged their mistake and ever since, in the field of psychotherapeutic research, the classic approach has rather undeservedly been known as Jacobson's RC.

Since RCIs are applied in non-experimental settings, researchers and methodologists have felt the need to correct for well-known effects that may disturb the assessment of change, including regression to the mean and practice effects. In this context, the attempt (RC_{ID}) of Hageman and Arrindell (1993) is particularly noteworthy. In fact the latter authors chose to take Kelley's estimate (Kelley, 1947) of the true difference score Δ_i of person i:

$$\Delta_i = \rho_{DD} D_i + (1 - \rho_{DD}) M_D \tag{1}$$

as the numerator of their index (where D_i is the observed difference of person i and M_D the mean observed difference in the research group). They used the standard error of measurement of the difference as the denominator, which is statistically justified in the classic approach but not in combination with Kelley's formula. Shortly after publication, this mistake was noticed by Zegers and Hafkenscheid (1994), who also proposed to this journal an improved index. This index, christened RC_{URCI}, also had Kelley's formula in the numerator but in the denominator it contained the following standard error:

$$\sigma_{\Delta.D} = \sigma_\Delta \sqrt{1-\rho_{DD}} = \sigma_D \sqrt{\rho_{DD}} \sqrt{1-\rho_{DD}} = \sigma_{E_D} \sqrt{\rho_{DD}}. \tag{2}$$

However, their proposal did not survive the review process. In early 1997, I offered to this journal a paper that demonstrates that Hageman and Arrindell's combination of Kelley's formula and their denominator was statistically not justified. This text was not clear enough to convince the reviewers, although one of them recognized Hageman and Arrindell's mistake too. A developed version of the paper, which also includes a discussion of Zegers and Hafkenscheid's RC_{URCI}, (Maassen, 2000a) is now being published by *Psychometrika*. So far, however, Zegers and Hafkenscheid have been less fortunate. Their RCI has only been mentioned in an official publication, as the starting point of the introduction of another new RCI (Bruggemans, Van de Vijver, & Huysmans, 1997).

In a recent publication, in which they appear to admit their earlier error, Hageman and Arrindell (1999) propose a new index RC_{INDIV}. They state: "RC_{INDIV} may also be considered an improved version of the RC_{ID} index (…). Though under standard conditions, RC_{ID} could be considered superior to RC in terms of correct classification of individuals, the present authors now recommend its even more precise successor RC_{INDIV}." The careful reader may have noticed that RC_{INDIV} is in fact identical to Zegers and Hafkenscheid's RC_{URCI}!, to which the condition $r_{DD} \geq 0.40$ has been added. Does this mean that this 'new' index is now statistically impeccable?

Unfortunately, it is not. In a subsequent publication (Maassen, 2000b), I demonstrate that the core of the confusion is caused by the fact that statistical testing and statistical estimation are not distinguished. The classic approach (i.e. Jacobson's RC) is a null hypothesis method. It uses a test statistic which has a standardized normal distribution under the null hypothesis of no true change. How to derive a *confidence interval* from such a statistic is a matter of basic statistics.

This interval constitutes the classic procedure. Zegers and Hafkenscheid's RC_{URCI} and thus Hageman and Arrindell's RC_{INDIV}, however, are based on estimation. Kelley's formula is a useful (but not the only and certainly not the most precise) estimate of a person's true change, using (1) his or her observed change and (2) the mean change in the sample. Expression 2 has been known for some time as the *standard error of estimate* belonging to Kelley's formula (McNemar, 1958; Lord & Novick, 1968, p. 67). This means that the true value of the difference of person *i* is comprised with probability 0.95 by the *estimation interval*:

$$\rho_{DD}D_i+(1-\rho_{DD})M_D\pm 1.96\sigma_{\Delta.D}.$$

The most salient problem that ensues from mixing up confidence and estimation intervals shows itself in the limit case $\rho_{DD}=0$. The reader will notice that no difficulty is encountered in the estimation procedure. (If $\rho_{DD}=0$, then also $\sigma_{\Delta.D}=0$, and the mean difference observed in the sample reveals itself as the best estimate. It is understood that this estimate is useless, because $\rho_{DD}=0$ implies that the variance of the true differences is zero and that with the pretest and posttest only errors are measured; then, the researcher is unable to assess changes.) However, when expression 2 is taken as the denominator of a RCI, the problem is obvious. Hageman and Arrindell try to protect their index against this danger by conditioning $\rho_{DD}\geq 0.40$. This condition is rather artificial, and surprising, in the context of difference scores which easily (but of course not necessarily) have a low reliability. It is also surprising in the light of their citation (on p.1174) of Rogosa, Brandt and Zimowski (1982): "the difference score can be an accurate and useful measure of individual change even in situations where the reliability is low".

A more serious problem in the context of psychotherapeutic research results from the fact that Kelley's estimate is not an unbiased estimate of a person's true change. It can be demonstrated that the bias is larger for persons who occupy an atypical position within the population to which they belong, for instance clients whose test scores indicate a strong deterioration during the therapy. Using the RCIs recently proposed, it might easily be concluded that such a person has significantly improved. Such erroneous results may well be obscured by the finding that overall the classification results of RC_{INDIV} and RC are much alike.

Although Hageman and Arrindell present a different argument for introducing RC_{INDIV} (i.e. maximization of Cronbach and Gleser's ϕ), it is instructive to interpret it as a null hypothesis method. It is then comparable with the classic approach and other shortcomings are revealed. Their use of the limits ±1.65 suggests that, under the null hypothesis of no true change, a standardized normal distribution is assumed for RC_{INDIV}. However, to transform Kelley's estimate into a *z*-score (a) its expected value should be subtracted, and (b) the outcome should be divided by its standard error, while sample information should be appropriately treated as stochastic. It can be shown (Maassen, 2000a) that this yields the following statistic:

$$\frac{\rho_{DD}D_i+(1-\rho_{DD})(M_D-\mu_\Delta)}{\sigma_{E_D}\sqrt{\rho_{DD}^2+\frac{(1+2\rho_{DD})(1-\rho_{DD})}{n}}}, \tag{3}$$

where μ_Δ is the population mean of the true difference scores. This expression is the correct expansion of Kelley's formula into a RCI. However, this formula is of theoretical interest rather than practical importance. The information about the population from which the research group

was selected is usually not available, forcing the researcher to use a large sample estimate. But when $n \rightarrow \infty$, the mean observed sample difference and the population mean of the true differences are approximately equal and expression 3 boils down to the classic approach! Thus, the classic approach can be considered as a large sample approximation of a properly constructed RCI based on Kelley's formula. If the research group can be considered as the population, the classic approach and the approach based on Kelley's estimate in fact become identical.

Speer (1999) fears that, in the future, a new proposal for a RCI will be published every 2–3 years. I am not sure that he is making a serious remark, but his fear may very well prove to be realistic, since researchers constantly feel the need to correct for various effects that endanger the internal validity of the assessment of change. As shown above, reviewers are sometimes more inclined to welcome incorrect new RCIs than to admit corrections of biased proposals already published. In particular, given the fact that the classic approach may prevent vulnerable clients of therapies from being falsely told their condition has improved, I strongly endorse Speer's remarks cited in the first paragraph.

References

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.

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.

Hageman, W. J. J. M., & 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.

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.

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., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maassen, G. H. (2000a). Kelley's formula as a basis for the asessment of reliable change. *Psychometrika,* (in press).

Maassen, G. H. (2000b). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology*, in press.

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

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). 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. (1999). What is the role of two-wave designs in clinical research? Comment on Hageman and Arrindell. *Behaviour Research and Therapy*, *37*, 1203–1210.

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.