Journal of the International Neuropsychological Society (2004), **10**, 899–901.
Copyright © 2004 INS. Published by Cambridge University Press. Printed in the USA.
DOI: 10.1017/S1355617704106115

DIALOGUE

Standard error in the Jacobson and Truax Reliable Change Index: The "classical approach" leads to poor estimates

NANCY R. TEMKIN, PHD
Departments of Neurological Surgery, Biostatistics, and Rehabilitation Medicine, University of Washington, Seattle

(Received April 21, 2004; Accepted May 7, 2004)

Abstract

Different authors have used different estimates of variability in the denominator of the Reliable Change Index (RCI). Maassen attempts to clarify some of the differences and the assumptions underlying them. In particular he compares the 'classical' approach using an estimate S_{Ed} supposedly based on measurement error alone with an estimate S_{Diff} based on the variability of observed differences in a population that should have no true change. Maassen concludes that not only is S_{Ed} based on classical theory, but it properly estimates variability due to measurement error and practice effect while S_{Diff} overestimates variability by accounting twice for the variability due to practice. Simulations show Maassen to be wrong on both accounts. With an error rate nominally set to 10%, RCI estimates using S_{Diff} wrongly declare change in 10.4% and 9.4% of simulated cases without true change while estimates using S_{Ed} wrongly declare change in 17.5% and 12.3% of the simulated cases ($p < .000000001$ and $p < .008$, respectively). In the simulation that separates measurement error and practice effects, S_{Ed} estimates the variability of change due to measurement error to be .34, when the true variability due to measurement error was .014. Neuropsychologists should not use S_{Ed} in the denominator of the RCI. (*JINS*, 2004, *10*, 899–901.)

Keywords: Test–retest data, Practice effects, Reliable change, Standard error of difference scores, Standard error of measurement of difference scores

Maassen (this issue, pp. 888–893) has taken a 'classical approach' to clarify the differences among some suggested estimators of variability for the Reliable Change Index (RCI) and its extensions. This classic approach, however, considers a practice effect to be a true change. Neuropsychologists are looking for change in an underlying condition and do not want to have 'normal' practice effects called a true change. Maassen extends the classical approach and begins to nicely lay out some of the differences in methods for theoreticians. Problems arise when these theoretical ideas are translated into practice. In my opinion, Maassen does a disservice to the practicing neuropsychologist by presenting a method that looks advantageous, but in fact, performs poorly when applied in the real world. Maassen advocates use of an estimator S_{Ed} which is supposedly based on measurement error alone. It is appealing because it yields smaller estimates of variability, and hence more sensitivity to true change, than the estimator used in our paper (Temkin et al., 1999). The estimator we used, S_{Diff}, is based on the variability of observed differences in scores in a group whose condition is not changing. As Maassen notes, the differences between the estimates is trivially small when the standard deviations at the two testings are the same. Maassen states, "The RCI of Temkin et al. accounts twice for differential practice effects." This is not true. As we show in simulations described below, when the standard deviations differ, Maassen's estimator indicates change substantially more often than it should when there is no true change, while the estimator we used maintains its nominal error rate. We feel that the small benefit of increased sensitivity to true change when the standard deviations are equal is far outweighed by the poor specificity of S_{Ed} when the standard deviations differ: That is, use of S_{Ed} when the standard deviations differ leads to greatly inflated rates of declaring either deterioration or improvement in people who have not truly changed.

Looking in more detail, Maassen starts from assumptions that seem questionable with respect to the findings one is likely to observe in longitudinal neuropsychological evaluations. They are classic assumptions that may be useful conceptually or theoretically, but which lead to methods that have undesirable properties in the real world. One problematic assumption is that somehow you know the true prac-

Reprint requests to: Nancy R. Temkin, Ph.D., Professor, University of Washington, Box 359924, 325 Ninth Avenue, Seattle, WA 98104-2499. E-mail: temkin@u.washington.edu

tice effect *in each individual* (i.e., as separate from the true change—which you are trying to assess—and from measurement error), and thus, any variation in practice effect from person to person is not a relevant part of variability when assessing change in an individual. This assumption is made explicit around Equation (1) in Maassen's paper, for example, "The practice effect in person i appears only in the numerator, assuming it to be fixed." In applying the RCI following Maassen's formulation, one subtracts the individual's known practice effect from their observed change in scores and compares the difference to S_{Ed}, which Maassen indicated represents the variability due to measurement error. It is difficult to think of a situation in the real world where one would know without error an individual's true practice effect but not also know their true change. Thus, if one is using the results in the real world, one is generally assuming a practice effect value of zero (original Jacobson and Truax RCI; Jacobson & Truax, 1991) or the mean in the reference population (Chelune et al., 1993) or some other approximation for the actual true practice effect in this individual.

How do the two estimators work in this more realistic situation? To evaluate the performance of the two methods, I did some simulations[1]. Starting with differences in variability seen in the Tactual Performance Test total min/block (TPTtotal), I simulated a reference sample of 500 observations from a normal distribution with means .52 and .43, standard deviations .49 and .33, and correlation .83, the values observed for TPTtotal in our reference sample (Dikmen et al., 1999). The estimated width of the RCI interval is .94 using S_{Diff} and .79 using S_{Ed}, a substantial difference as pointed out by Maassen. To test how the different intervals perform, I simulated another 1,000 cases from the same distribution. After subtracting off the mean difference in the reference sample, 104 of these 1,000 'new cases' were wrongly declared to have improved or deteriorated when the interval was based on S_{Diff}. This is almost identical to the 100 cases one would expect by using a 90% interval. When the interval was based on S_{Ed}, however, 175 of the 'new cases' were wrongly declared to have changed. This is 75% higher than the 100 that one would expect and highly significantly different from the nominal 10% ($p <$ .000000001). Thus we see that rather than the interval based on S_{Diff} overestimating the variability, it accurately accounts for the variability while the interval based on S_{Ed} underestimates the variability, making the interval too short and the error rate too high. Note that there was no explicit differential practice effect in this simulation—just bivariate normal observations with different standard deviations.

To provide further information about the effect of practice, I tried to simulate 'true' values for an individual, independent measurement error at the two times and independent practice effects with the resulting sums (*'true'* + *error1*, *'true'* + *practice* + *error2*) having a bivariate normal distribution as observed for TPTtotal. I couldn't do it. When forming sums of independent effects (as Maassen describes for the classical approach), the correlation can be no more than the ratio of the smaller to the larger standard deviation (Wilks, 1962). Thus, for TPTtotal, that is .33/.49 or .67. The maximum occurs when there is no measurement error. To come close to the distribution of TPTtotal, I simulated 'true' values with a mean of .43 and standard deviation of .32, errors with mean zero and standard deviation .01, and practice with mean .09 and standard deviation .37. This yields sums with mean .43 and standard deviation .32 for one time and mean .52 and standard deviation .48 for the other with correlation of .65. Carrying out a similar investigation, from a reference population of 500 cases, the interval width was 1.20 based on S_{Diff} and 1.10 based on S_{Ed}—a smaller but still nontrivial difference, as one would expect since the observations are less highly correlated than was actually observed. The error rates tell the same story, however. When basing the intervals on S_{Diff}, 94 of the next 1000 'new cases' were declared to have improved or deteriorated, very near the nominal 10%. When basing the intervals on S_{Ed}, 123 of these same 'new cases' were declared to have improved or deteriorated, an error rate 23% higher and significantly greater than the nominal 10% ($p < .008$). Now in this case, we actually know each individual's practice effect (since we simulated the data). If we subtract it from the change to calculate the RCI as Maassen suggests, does S_{Ed} then have the nominal 10% error rate? Definitely not! In the 1000 'new cases,' not a single one was declared to have changed. In fact, the largest of the 1000 observed differences was under .08, while it would need to be over .55 to be called changed based on S_{Ed}. Indeed, S_{Ed} equals .34, while the true variability due to measurement error yields a standard deviation of .014. Thus, while S_{Diff} is appropriately accounting for variability due to measurement error and variations in practice effect, S_{Ed} is not appropriately accounting for either. It may work in theory, but when the correlation is estimated from data with variable practice effect (or any other reason that causes unequal variability), S_{Ed} as defined and estimated by Maassen (this issue), loses both its clear interpretability and its stated error rate.

Maassen's paper (this issue) also may suggest to practitioners that one can base an RCI on estimates of variability and/or reliability that come from different samples or samples where there may be true change in some individuals. This is likely to lead to a very poor estimate of the variability of the RCI. Maassen's paper seems to be misinterpreting the population we used in Temkin et al. (1999) and by implication what we recommend. He states that we "have interpreted Expression 5 as referring to the standard deviation of the *observed difference scores in the research group at hand*" (italics the author's). We have instead interpreted it as the standard deviation of the *observed difference scores in a group of individuals for whom we would expect no true*

[1] Simulations and significance tests were done using Microsoft Excel (Microsoft Corp., 1999). For the significance tests, the p value was obtained as the probability that a binomial distribution with 1000 trials each having probability .10 would yield at least this many 'successes.' In this case, a 'success' would be erroneously declaring an observed change to be a deterioration or improvement.

https://doi.org/10.1017/S1355617704106115

change. The difference is critical. If one had a group of people with mild cognitive impairment and wanted to decide whether the performance of some had deteriorated, using the standard deviation of the difference scores in the group as a whole when some likely have true change would clearly overestimate the variability. The groups included in the Temkin et al. (1999) paper (normals in a clinical study, friends of people with head injury, people hospitalized for trauma that spared the head) were chosen specifically because a true change was not expected. What you see in the differences in such a population is the normal variability from testing to testing in real people, whether that variability comes from unreliability of the instrument (measurement error) or unreliability of normal people (which some might call differential practice). I would call such a population an *external source*, as both Maassen and we advocate using. Our original paper might cause some confusion because we did indeed use the same population to calculate the percent of cases outside the limits calculated by the different methods. This was not to identify which of these normal individuals has true change, but to see the effects of practice (with Method 1) and nonnormality (with all methods) on misclassification rates in a sample where true change can be assumed to be nonexistent. Note that use of a sample that includes people with true change to estimate the standard deviations and correlation, as Maassen seems to be suggesting just before he presents the estimate of S_{Ed} (Expression 6), is going to inflate the variability for the RCI estimated by S_{Ed} just as it will for S_{Diff}. It may be even worse to take a published estimate of the reliability coefficient and combine that with study-sample-based estimates of standard deviation. A correlation coefficient is strongly influenced by the spread of the true values. If the spread of true values in the sample used to estimate the reliability coefficient differs from that in the research group, using that reliability coefficient with the observed standard deviations in any of the estimators will not give an accurate estimate of the variability for the RCI.

Maassen's discussion of regression to and from the mean is also apt to confuse some readers. Maassen takes a term—regression to the mean—that commonly refers to a phenomenon resulting from independent measurement error, and uses it for a result of differential practice effects that relate linearly to the initial score. Maassen writes an equation for such a relationship and, assuming that equation to be true, he calculates the slope β that would yield the differences in variability observed. It should not be surprising that "the posttest variance exceeds the pretest variance in every instance where $\beta > 1$" (Maassen, this issue). It would be shocking if it were otherwise. As pointed out by Maassen, $\beta = b/\rho_{xx}$ (attributed to McNemar, 1969, p. 173). To obtain an estimate of β, Maassen substitutes r_{xy} for ρ_{xx}. However, $b = (S_y/S_x)r_{xy}$ (Draper & Smith, 1966, p. 35), so the estimate of β equals $b/r_{xy} = S_y/S_x$, which will exceed 1 if and only if the posttest variance exceeds the pretest variance. It seems unlikely that this is going to help many readers to better understand the different ways of estimating the variability for the Reliable Change Index. In fact, by using a common term differently from its usual definition, Maassen may be adding unnecessary confusion.

In summary, Maassen attempts to provide some clarification about the differences in the estimates of standard deviation that have been used in variants of the Reliable Change Index. However, he provides poor guidance to practitioners about what to use if they want to decide whether there is evidence a patient has exhibited real change between two testings. S_{Ed} provides only a trivial increase in sensitivity when the variability at the two testings is the same. When the variability of the two testings differ, the increased sensitivity provided by use of S_{Ed} can appear to be substantial, but this is exactly the situation when use of S_{Ed} leads to what would probably be considered unacceptable specificity. For a practicing neuropsychologist wanting an RCI-type assessment, the trade-off favors using S_{diff} estimated from a single sample where true change can be ruled out by the circumstances. S_{diff} can be estimated most accurately directly from the individual differences, but it can also be estimated using Maassen's Expression 5 with all components coming from the same sample.

ACKNOWLEDGMENTS

This work was supported by the following grants: NIH-NINDS R01 NS19463, NIH-NICHD U01 HD42653.

REFERENCES

Chelune, G., 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.

Dikmen, S.S., Heaton, R.K., Grant, I., & Temkin, N.R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, *5*, 346–356.

Draper, N.R. & Smith, H. (1966). *Applied regression analysis*. New York: John Wiley & Sons, Inc.

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

Maassen, G.H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society*, *10*, 888–893 (this issue).

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

Microsoft Corp. (1999). Excel 2000 [Computer software]. Redmond, WA: Author.

Temkin, N.R., Heaton, R.K., Grant, I., & Dikmen, S.S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*, 357–369.

Wilks, S.S. (1962). *Mathematical statistics*. New York: John Wiley and Sons, Inc.

https://doi.org/10.1017/S1355617704106115