Journal of the International Neuropsychological Society (2004), **10**, 894–898.

DOI: 10.1017/S1355617704106103

DIALOGUE

Holding out for a reliable change from confusion to a solution: A comment on Maassen's "The standard error in the Jacobson and Truax Reliable Change Index."

ANTON D. HINTON-BAYRE

Cognitive Psychophysiology Laboratory, School of Medicine, The University of Queensland, Australia

(Received April 20, 2004; Accepted May 7, 2004)

It is important to preface this piece by advising the reader that the author is not writing from the point of view of a statistician, but rather that of a user of reliable change. The author was invited to comment following the publication of an original inquiry concerning Reliable Change Index (RCI) formulae (Hinton-Bayre, 2000) and after acting as a reviewer for the current Maassen paper (this issue, pp. 888–893). Having been a bystander in the development of various RCI methods, this comment serves to represent the struggle of a non-statistician to understand the relevant statistical issues and apply them to clinical decisions. When I first stumbled across the 'classical' RCI attributed to Jacobson and Truax (1991) (Maassen, this issue, Equation 4), I was quite excited and immediately applied the formula to my own data (Hinton-Bayre et al., 1999). Later, upon reading the Temkin et al. (1999) paper I commented on what seemed to be an inconsistency in their calculation of the error term (Hinton-Bayre, 2000). My "confusion" as Maassen suggests was derived from the fact that I noted the error term used was based on the standard deviation of the difference scores (Maassen, Expression 5*) rather than the Jacobson and Truax formula (Maassen, Expression 4). This apparent anomaly was subsequently addressed when Temkin et al. (2000) explained they had employed the error term proposed by Christensen and Mendoza (1986) (Maassen, Expression 5). My concern with the Maassen manuscript was that it initially appeared two separate values could be derived through using expressions 5 and 5* using the Temkin et al. (1999) data. This suggested there might be four (expressions 4, 5, 5*, and 6), rather than three, ways to calculate the reliable change error term based on a null hypothesis model. Once again I was confused. Only very recently did I discover that expressions 5 and 5* yield identical results when applied to the same data set (N.R. Temkin, personal communication) and when estimated variances are used (G. Maassen, personal communication). The reason for expressions 5 and 5* yielding slightly different error term values using the Temkin et al. (1999) data was due to use of nonidentical samples for parameter estimation. The use of non-identical samples came to light in the review process of the present Maassen paper—which Maassen now indicates in an author's note. Thus there were indeed only three approaches to consider (Expressions 4, 5, & 6). Nonetheless, Maassen maintains (personal communication) that Expression 5, as elaborated by Christensen and Mendoza (1986), represents random errors comprising the error distribution of a given person, whereas Expression 5* refers to the error distribution of a given sample. While it seems clear on the surface that the expressions represent separate statistical entities, it remains unclear to the present author how these expressions can then yield identical values when applied to test–retest data derived from a single normative group. Unfortunately however, my confusion does not stop there.

It is readily appreciable that the RCI_{JT} (Expression 4) is relevant when only pretest data and a reliability estimate are available and no true change is expected (including no practice effect). When pre- and posttest data are available in the form of test–retest normative data it seems sensible that posttest variance be included also. Expression 6 appears a neat and efficient method of incorporating posttest variance. And, according to Maassen, it remains so whether or not pre- and posttest variances are believed to be equivalent in the population (see also Abramson, 2000). Given that test–retest correlations will always be less than unity, if measurement error alone accounts for regression to the mean, then pre-and posttest variances should not differ (Maassen, personal communication). Maassen suggests that differ-

Reprint requests to: Anton D. Hinton-Bayre, Ph.D., Cognitive Psychophysiology Laboratory, School of Medicine, The University of Queensland, Herston Road, Herston, Queensland, Australia, 4064. E-mail: s309339@student.uq.edu.au

ences between pre- and posttest variances can be attributed to differential practice. This is explained through reference to fanspread and regression to the mean where a relationship (positive or negative) is seen between pretest scores and individual practice effects.

Expression 5 also appears to incorporate posttest variability. The two expressions differ in how they purport to account for the presence of a 'differential practice effect.' The differential practice effect is the extra variation added by the individual's true difference score (Δ_i) and their practice effect (Π_i)—see the expression following Expression 7 (Maassen, this issue). Temkin (this issue) appears to argue that the individual practice effect cannot be known, and thus differential practice effect is estimated in the numerator of the expression comparing pre- and posttest scores. Moreover, as differential practice is estimated it should be incorporated into the error term as provided by Expression 5. Maassen argues that an individual's posttest score is in part affected by an individual differential practice effect, thus this 'extra' element of variance is not required in the error term. Maassen asserts that it has already been taken into account through incorporating posttest variance. Temkin maintains that individual differential practice effects are excluded from the Maassen error term. It must be remembered that when pre- and posttest variances are equal, the two error estimates will be identical—there would be no differential practice effect according to Maassen.

The discrepancy between estimates becomes increasingly pronounced as the pre- and posttest variance estimates differ and as the reliability improves (see Maassen, Expression 7). Given that clinical decisions should not be made using results taken from unreliable measures, the present author sees the reliability component as a lesser concern in practice. Clinically one could argue that a reliability of $r > .90$ is a minimum for assisting in decisions regarding any individual's performance. Moreover, to derive RCI cut scores using measures with reliability estimates $r < .70$ will yield intervals so wide as to be clinically useless in many cases. The RCI should not be considered a correction for the use of an unreliable measure. Thus, clinically speaking, the emphasis rests more squarely on the differences between pre- and posttest variances and the subsequent effects on the error term under Expressions 5 and 6.

The above comments are not restricted to the use of Expression 6, but also apply, at least in part, to Expression 5. The Temkin et al. (1999) approach obviously affords greater protection against type 1 errors in decision-making (false positives). To what extent the adjustment is valid is still obviously under conjecture. There is an obvious clinical appeal to evaluating whether the difference score obtained by a person suspected of change is 'unusual' when compared to a normative set of difference scores from those not expected to change (Expression 5*). This is akin to determining impairment *via* simple standard (*Z* or T) scores as commonly done with neuropsychological tests. Yet, Temkin and colleagues' approach might be criticized as a departure from the accepted standard error of measurement approach that Maassen continues to advocate.

Maassen suggests that Expression 6 yields a better estimate than Expression 5 in any circumstance, and illustrated the discrepancy between error terms for McSweeny et al.'s (1993) PIQ data. This comparison seems fair, as the distribution of scores was presumably normal. What is not known is how did this 8% difference in the magnitude of error affect the false positive rate in real terms. A point investigated more directly below. The 25% discrepancy in magnitude of error noted between expressions on Temkin et al.'s (1999) TPT total scores are of lesser interest. Given that neither method will yield acceptable false positive rates when data is obviously skewed in a large sample it would seem prudent to operate on distribution-free based intervals.

The evolution of various forms of RCI has been a challenge to follow. The interested clinician has been forced to grapple with a multitude of parameter-based expressions attempting to best account for error in its various forms. If there is ever a need for the bridge between academia and clinical practice to be built, then this is surely one of those occasions. Ultimately the clinician needs a method that reliably and validly allows for an interpretation of individual change. The question remains however, which one do we choose? This was when I realized my confusion had still not abated. Adopting a statistically conservative approach one might employ the *sometimes* wider Expression 5 estimate, yet actual change might be missed. This could occur when making a before-after comparison as seen in the case of brain surgery, or when wishing to plot recovery in an already affected group, such as following progress after brain injury. But of course this must be balanced against the possible decision errors stemming from the use of the *sometimes* narrower Expression 6 estimate. Clearly, this is not an issue that should be resolved solely through reference to the relative importance of decision errors.

Thus, in comparing the two approaches for estimating the RCI error term, one should consider psychometric, statistical, applied, and clinical elements. On a psychometric level, the difference appears to depend on whether the error variance should reflect what is unusual for that person (Expression 6) or whether that person is unusual (Expression 5*). Classical psychometric theory would suggest the concept represented by Expression 6 is preferable. However, the present author still cannot see how if Expressions 5 and 5* yield the same value when derived from a single group retested (as would be done using retest norms), that Expression 5 (and thus 5*) does not also represent an estimate of individual error. On a statistical level, the difference between methods appears to hinge on the management of differential practice in the error term. It seems reasonable that Expression 6 would be preferred when no differential practice or true change was present. Yet, one must remember that under such circumstances the two expressions will agree. Moreover, it does seem clear that one cannot always readily determine whether differential practice is present, as will be demonstrated below. On an applied

https://doi.org/10.1017/S1355617704106103 Published online by Cambridge University Press

level, the question becomes which method, if either, is desirable when systematic practice effects are known to occur. It stands to reason also that if pretest scores can predict practice (or posttest scores) so too can other factors. Further on this point, if one refers to Temkin et al. (1999) Table 5 (p. 364), it can be seen that the error terms for the regression based models—both simple and multiple—are narrower than the Maassen prediction intervals ($S_{ED} \times 1.645$) for Expression 6 (excepting simple linear regression for PIQ). *Prima facie* this seems to suggest that more than just pretest scores may be relevant to determining posttest scores or practice. Maassen himself refers to other works when discussing how to deal with practice effects using either mean adjustments (Chelune et al., 1993) or regression based reliable change models (McSweeny et al., 1993; Temkin et al., 1999). It is not clear how the proposed adjustment to error made by Expression 6, compares conceptually to the seemingly more versatile and efficient regression-based models. This is an important point as practice effects are a frequent concern when conducting repeated testing on performance based measures that are commonly used in neuropsychological assessment. To this end, it is unclear whether the present Maassen error term (Expression 6) is preferred in those frequent instances where systematic practice effects are at work. Finally, and arguably most importantly, on a clinical level, it is worth considering whether the use of one method over another yields demonstrable differences in false positive rates. To further address the latter issue, Expressions 5 and 6 were applied to previously published data to examine the rate of significant change on retesting in a sample where change was not expected.

A normative series of 43 young male athletes were tested twice (preseason) at a retest interval of 7–14 days on the written versions of the Digit Symbol Substitution test of the Wechsler scales, the Symbol Digit Modalities test, and the Speed of Comprehension test (Hinton-Bayre et al., 1997). These data were fortuitous as across the three measures reliability estimates were all $r > .70$, yet the magnitude of the difference between test and retest variances was not consistent (see Table 1), thereby providing an extra example of how such differences may affect error and subsequent classification of change. Significant mean practice effects were found on Digit Symbol and Speed of Comprehension, but not the Symbol Digit, using repeated measures *t* tests (see Table 1). Variance estimates were equivalent on Digit Symbol, with an estimated disattenuated regression coefficient (*est.* β_C) approximating one. Maassen indicated $\beta_C = b_c/\rho_{xx}$, in this instance β_c was estimated where r_{xy} was substituted for ρ_{xx} as a measure of pretest reliability. Not surprisingly, the error terms derived for Expressions 5 and 6 differed only at the fourth decimal place. When using the RCI adjusting for mean practice (Chelune et al., 1993) an equal number of participants were classified as having changed significantly based on each expression (see Table 1), at rates consistent with chance using a 90% level of confidence (overall rate 11.6%). Pretest scores were found to correlate negatively with difference scores (posttest − pretest) suggesting regression to the mean. On the Symbol Digit, the standard deviations for Times 1 and 2 were more discrepant, with posttest variance less than pretest variance (thus *est.* $\beta_C < 1$). A stronger negative correlation was seen between pretest and difference scores on Symbol Digit (see Table 1). Expression 5 produced an interval 3.6% larger than Expression 6 and classified 5 participants (11.6%) as significantly changing, whereas the marginally narrower Expression 6 identified 2 additional participants as having significantly improved (overall rate 16.3%). Yet both the extras only recorded $z = 1.649$, and thus barely reached significance (90% *C.I.* = ±1.645), and might have been overlooked if intermediate calculations were rounded and would possibly only be judged clinically changed in the context of other test changes. On the Speed of Comprehen-

Table 1. Test–retest data and reliable change estimates for a normative sample of male athletes ($N = 43$)

	DSS	SDM	SOC
Time 1, *M*(*SD*)	60.19 (12.77)	57.16 (12.56)	52.49 (16.48)
Time 2, *M*(*SD*)	66.42 (12.72)	59.58 (9.95)	63.47 (19.26)
RM *t* (*p*)	5.26 ($p < 0.001$)	1.86 ($p = 0.07$)	5.46 ($p < 0.001$)
$r_{X.Y}$	0.814	0.736	0.739
$r_{X.Y\text{-}X}$	−0.310*	−0.614*	−0.190
est. β_C	0.997	0.792	1.169
Expression 5 (S_{diff})	7.7670	8.5363	13.1828
Expression 6 (S_{ED})	7.7669	8.2369	12.9640
Regression (S_{EE})	7.4746	6.8188	13.1460
Improved ($S_{diff}/S_{ED}/S_{EE}$)[a]	2/2/2	2/4/3	3/3/3
Deteriorated ($S_{diff}/S_{ED}/S_{EE}$) [a]	3/3/2	3/3/4	2/2/2

Note. SDM=Symbol Digit Modalities, DSS=Digit Symbol Substitution, SOC=Speed of Comprehension. RM *t* = repeated measures *t* test, $r_{X.Y}$ =test–retest correlation, $r_{X.Y-X}$ = correlation between pretest and difference scores, *est.* β_C =estimated disattenuated regression coefficient.
[a]Number of participants changed based on 90% confidence interval, adjusting for practice.
*$p < .05$ (two-tailed)

https://doi.org/10.1017/S1355617704106103 Published online by Cambridge University Press

sion, posttest variance was larger than pretest variance (thus *est*. $\beta_C > 1$). In this instance Expression 5 error was only 1.7% larger than Expression 6 error. Again classification rates were equal for both expressions and consistent with chance. However, the correlation between pretest and difference scores failed to reach significance (see Table 1).

In summary, the error terms derived from Expressions 5 and 6 in a new set of data were comparable in magnitude, and similar false positive classification rates were observed. The only classification discrepancy was seen on the Symbol Digit, where the difference between pre- and posttest variances was greatest. The apparent increase in false positives for Expression 6 on the Symbol Digit should not be overstated given the small sample size and borderline nature of the additional cases noted as changing. Replications contrasting the two methods using large sets of real data would be more convincing that any clinical differences observed here are more real than apparent.

It was observed that, as pretest variance exceeded posttest variance, the relationship between pretest and difference scores became increasingly negative. In other words, relatively greater pretest variance was linked to more obvious regression to the mean. Maassen stated, "If $\beta_C = 1$, there is no better estimation for all the practice effects in the normative population than the population mean (for practice) π_c," or that differential practice effects are not present. When *est*. β_C approximated 1 as was seen with the Digit Symbol, there was still a significant negative relationship between pretest and difference scores, suggesting regression to the mean should be due to measurement error. It was also noted that when *est*. $\beta_C < 1$, regression to the mean was greater still, suggesting regression to the mean due to differential practice effects and measurement error. Moreover, when *est*. $\beta_C > 1$, neither regression to the mean nor fanspread was observed. These findings may reflect the substitution of r_{xy} for ρ_{xx}, but no other estimate for pretest reliability is practically available for most timed performance measures. Maassen (personal communication) has suggested that as r_{xy} is a lower-bound estimate of ρ_{xx}, the values derived for *est*. β_C will be overestimates. For example, the true value of β_C on Digit Symbol would be less than 1, suggesting the effects of differential practice on top of measurement error in producing regression to the mean. On the Symbol Digit, β_C would be further below one suggesting a stronger role of differential practice on regression to the mean for that measure. The true Speed of Comprehension β_C would more closely approximate one and thus differential practice would not be expected, thus making it less likely to see a relationship between pretest and difference scores. These data support the notion that regression to the mean affects posttest scores and thus needs to be taken into account. It also supports the understanding of independence between differential practice and measurement error as contributors to regression to the mean. The difficulty arises from the realization that as ρ_{xx} cannot be readily obtained in most instances for performance based measures. It is subsequently difficult to estimate β_C and thus determine whether differential practice effects exist. In this way, the relative contributors of any regression to the mean cannot be determined. As for which of Expressions 5 and 6 best achieves protection against regression to the mean when systematic practice effects exist may be moot, given a better understanding of regression-based RCI.

Regression based analyses were conducted for the sake of comparison, given the presence of mean practice and possible differential practice effects observed in the measures. RC scores were calculated using the following formulae based on McSweeny et al. (1993) and Temkin et al. (1999):

$$Y' = bX + a$$

$$S_{EE} = S_Y\sqrt{1 - r^2}$$

$$RCI = \frac{Y - Y'}{S_{EE}},$$

where X = pretest score, Y = actual posttest score, Y' = predicted posttest score, b = slope of the least squares regression line of Y on X, a = intercept, S_{EE} = standard error of estimate, S_Y = estimated standard deviation of posttest scores, and r^2 = the squared correlation coefficient.

Not surprisingly, when regression to the mean was greatest (Symbol Digit), the regression-based error term was considerably smaller than the error terms from either Expression 5 or 6 (see Table 1). The proportional reduction in error seen when variances were equal (Digit Symbol) was much less pronounced. Further, the S_{EE} was not the smallest error estimate when the regression to the mean was not significant (Speed of Comprehension). Interestingly, the classification rates based on the regression model with S_{EE} were not remarkably different than rates seen with Expressions 5 and 6. It was noted that the relatively higher number of false positives on Symbol Digit using the S_{EE} were seen in participants with more extreme pre- or posttest scores compared to the rest of the group. The regression model accounts for regression to the mean when present and subsequently provides a smaller error estimate. However, it would appear that the simple regression based model (in its current form) might not necessarily provide the most efficient error term when regression to the mean is not clearly evident, nor improve false positive rates. A multiple regression model was not investigated here given the small sample, and subsequent instability of predictors.

It must be noted that the results presented above focus more on apparent trends and may possibly reflect idiosyncrasies of the data. Nonetheless, the results further suggest that the practical difference between the error term values obtained from Expressions 5 and 6 will most often be negligible. This reflects data presented by Temkin et al. (1999) and McSweeny et al. (1993). This is not to imply that either expression will suffice, as ultimately the best estimate of error should be used in any instance. Concordantly, rather

https://doi.org/10.1017/S1355617704106103 Published online by Cambridge University Press

than endorsing either Expression 5 or 6, the present author awaits further consideration of the regression based models, particularly in comparison to the approaches discussed here when practice effects are evident (e.g., Maassen, 2003; Temkin et al., 1999).

ACKNOWLEDGMENTS

The author would like to thank Karleigh Kwapil, Dr. Nancy Temkin and Dr. Gerard Maassen for their comments on drafts of this manuscript.

REFERENCES

Abramson, I.S. (2000). Reliable change formula query: A statisticians comment. *Journal of the International Neuropsychological Society*, *6*, 365.

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 single subject: An alteration of the RC Index. *Behavior Therapy*, *17*, 305–308.

Hinton-Bayre, A.D. (2000). Reliable change query. *Journal of the International Neuropsychological Society*, *6*, 362–363.

Hinton-Bayre, A.D., Geffen, G.M., Geffen, L.B., McFarland, K.A., & Friis, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. *Journal of Clinical and Experimental Neuropsychology*, *21*, 70–86.

Hinton-Bayre, A.D., Geffen, G., & McFarland, K. (1997). Mild head injury and speed of information processing: A prospective study of professional rugby league players. *Journal of Clinical and Experimental Neuropsychology*, *19*, 275–289.

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. (2003). *Principes voor de definitie van reliable change (2): reliable change indices en practice effects* [Principles of defining reliable change (2): Reliable change indices and practice effects]. *Nederlands Tijdschrift voor de Psychologie*, *58*, 69–79.

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).

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. *Clinical Neuropsychologist*, *7*, 300–312.

Temkin, N.R. (2004). Standard error in the Jacobson and Truax reliable change index: The 'classical approach' leads to poor estimates. *Journal of the International Neuropsychological Society*, *10*, 899–901 (this issue).

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

Temkin, N.R., Heaton, R.K., Grant, I., & Dikmen, S.S. (2000). Reliable change query: Temkin et al. reply. *Journal of the International Neuropsychological Society*, *6*, 364.