REVIEW

Reliable change assessment in sport concussion research: a comment on the proposal and reviews of Collie *et al*

G H Maassen

Br J Sports Med 2005;**39**:483–488. doi: 10.1136/bjsm.2004.015594

The proposal of Collie *et al* for determining reliable change in sports medicine and their review of other current procedures did not address mathematical or theoretical aspects. This article attempts to fill this lacuna, in order to examine the validity of their proposal and the suitability of their review. The conclusions drawn are that Collie *et al* presented their method too carelessly and their review of other methods leaves serious drawbacks and mistakes unnoticed. A scheme is here provided showing which of the current methods can be used in which situation.

The major aim of the present review is to provide a substantive discussion of the reliable change indices (RCIs) encountered in sports medicine literature, in particular the proposal of Collie *et al*[1] for the standard error to be used. Since Hinton-Bayre *et al*[2] advocated the use of the RCI as an objective criterion in sport concussion management, several alternative proposals for the standard error have been presented. All these proposals[1–3] carry the names S_{diff} (or SD_{diff}), referring, however, to different formulas. The alternative standard errors were chosen without convincing substantive argument, which will leave the clinician in the field confused. The present article attempts to clarify the confusion by discussing their characteristics and underlying assumptions. Secondly, it comments on the review of other current procedures by Collie and colleagues.[4 5] The fact that I am here personally involved was an extra impetus for writing this review.

APPROACHES TO RELIABLE CHANGE ASSESSMENT IN SPORTS MEDICINE

Several decades ago, McNemar[6] proposed to assess the reliability of change in an individual i between two measurement moments by considering the observed difference score divided by the standard error of measurement of the difference score: D_i/σ_{ED}. When the absolute value of this statistic exceeds 1.96, the null hypothesis of zero true change is rejected and the change was called dependable by McNemar, as the term reliable change was not yet in use. McNemar also provided various formulas for calculation of the standard error under various assumptions within the population of which the person to be assessed is a member. For a general understanding, citing of the basic formula from which all these formulas are derived is inevitable:

$$\sigma^2_{E_D} = \sigma^2_{E_X} + \sigma^2_{E_Y} = \sigma^2_X(1-\rho_{XX}) + \sigma^2_Y(1-\rho_{YY}) \quad (1)$$

Application of this expression requires knowledge (preferably from external sources) of the standard errors of measurements of the initial testing X and the final testing Y of the population variances and the reliability coefficients of X and Y. The procedure and formulas are based on classical test theory (CTT), and this procedure was christened the classic approach to reliable change assessment.[7 8]

Hinton-Bayre *et al*[2] were the first to introduce the reliable change methodology in sport concussion research. They adopted the RCI developed by Jacobson and colleagues,[9–11] who were the first to use the term.[9] See my recent article[12], for a historical account of the development of the RCI of Jacobson and colleagues, which was eventually established in the psychotherapy literature by Jacobson and Truax[11], and which is, for that reason, also known as the JT index. The JT index comprises a standard error that is in fact equal to the square root of expression 1 adapted to the case where it is assumed that pretest and posttest are parallel—that is, equivalent alternative—measures having equal standard errors of measurement. The RCI of Jacobson *et al* was meant to be applied in psychotherapy research where it is plausible to assume the absence of confounding effects—for example, practice effects, referring to the phenomenon that individuals who take a test for a second time, or take an alternative form of the test, do better than those who take the test for the first time—thus making the assumption of parallel measures credible. Furthermore, Jacobson and Truax estimated the parameters in expression 1 from the values observed in a control group of healthy people—in particular, the common reliability coefficient of pretest and posttest was estimated by the test-retest reliability. Unfortunately, following Christensen and Mendoza,[13] the denominator was now carelessly named standard error of the difference score and denoted as S_{diff}. Thus the formula of the standard error proposed by Jacobson and Truax and adopted by Hinton-Bayre *et al* became:

$$S_{diff} = S_{E_X}\sqrt{2} = S_X\sqrt{2(1-r_{XY})} \quad (2)$$

Abbreviations: CTT, classical test theory; RCI, reliable change index

Correspondence to: Dr Maassen, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, PO Box 80. 140, Utrecht 3508 TC, the Netherlands; g.maassen@fss.uu.nl

Accepted 15 March 2005

Iverson and colleagues[3] adapted expression 1 in a different way. They only assumed that pretest and posttest have a common reliability coefficient, to be estimated by the test-retest reliability observed in a control group of healthy people. The other parameters in expression 1 are estimated by the values observed in this control group as well. Iverson *et al* also denoted their standard error as S_{diff}, but for sake of distinction, in the present context I prefer to use S_{ED}. Easy algebra then leads to the following expression for the standard error:

$$S_{E_D} = \sqrt{(S_X^2 + S_Y^2)(1 - r_{XY})} \qquad (3)$$

A third expression for the standard error was recently proposed by Collie *et al*[1]:

$$SD_{diff} = \sqrt{\frac{1}{N-1}\sum_k (D_k - \overline{D})^2} \qquad (4)$$

referring to the differences between paired observations observed in a control group of size N. Contrary to the preceding standard errors, this expression is not derived from expression 1, but basic statistics textbooks such as that by McNemar[6] demonstrate the following equation:

$$SD_{diff} = \sqrt{\frac{N}{N-1}(S_X^2 + S_Y^2 - 2S_X S_Y r_{XY})} \qquad (5)$$

This equation reveals a link between expression 4 and the careless naming (standard error of the difference instead of standard error of measurement of the difference) and the following ambiguous equations presented by Christensen and Mendoza[13]:

$$S_{diff} = \sqrt{S_X^2 + S_Y^2 - 2S_X S_Y r_{XY}} = \sqrt{2S_{E_X}^2} \qquad (6)$$

The left hand equation can be interpreted in two different ways: (*a*) referring to the stochastic pretest and posttest scores observed in a single person (in which case the expression in the middle should be named standard error of measurement of the difference) or (*b*) referring to the pretest and posttest scores observed within a sample of people (in which case the expression in the middle is justly named standard error of the difference). Interpretation (*a*) was clearly the intention of Christensen and Mendoza, as the right hand equation only holds in interpretation (*a*), under the assumptions of parallel measures and zero effects of testing, and under the null hypothesis of zero true change. Moreover, they described the RCI as a statistic that ''represents the amount of difference which one could expect between two scores obtained on the same test by the same individual as a function of measurement error alone'' (p 307).[13] However, authors unfamiliar with the classical test theory disregarded interpretation (*a*) and adhered to the conceptually wrong interpretation (*b*).[14 15]

Collie *et al*[1] cited correctly the formula of the standard error of Jacobson and Truax (although consistently misspelling the latter author's name). They also are obviously aware of the difference between the concepts of the standard error of the difference and the standard error of measurement of the difference, as they state, probably referring to earlier studies:[14 15] ''The problem with currently applied RC calculations is that the normal amount of variability in change over time within individuals is estimated on the basis of differences between individuals at a single time point! There is no reason to believe that variation between individuals at one time point accurately represents the variation within individuals between two time points.'' (p 371).[1] Nevertheless, they propose to calculate the standard error according to expression 4, stating that ''this alteration to previous RC calculations produces an RC calculation that is mathematically and theoretically correct''. Thus, after all, Collie *et al* followed interpretation (*b*). Contrary to the classic approach, where an individual's observed difference is judged by what is regarded exceptional within this individual's theoretical distribution of measurement errors, Collie *et al* judge an individual's observed difference score by what is designated exceptionally large within the normative sample.

A THEORETICAL COMPARISON OF THE DIFFERENT APPROACHES

The RCIs applied in sports medicine are composed of the same numerator—that is, the observed change in a given person—but different standard errors in the denominator. To what extent will the outcomes of these indices differ, and is there a preferred approach?

Expression 3 (the standard error used by Iverson *et al*) is clearly preferable to the standard error of the JT index in the case in which the initial and final variance really differ. However, when it is plausible to assume that the variance remains unchanged, the observed initial and final variance may still differ on the basis of chance. The common theoretical value should not be estimated by only the initial variance (as Jacobson and Truax proposed) but by pooling initial and final variance. The outcomes of the JT index and those of the RCI with expression 3 in the denominator will differ to the extent that the final variance differs from the initial variance.

Although it has already been noted that, for conceptual reasons, expression 3 is also preferable to expression 4, it is still worth examining the difference between outcomes of the two approaches. According to CTT, the variance of the differences observed in the control sample can be split into two components: a variance of ''true'' changes (including effects of testing) and an error variance component:

$$SD_{diff}^2 = Var(\Delta + \pi) + S_{E_D}^2 \qquad (7)$$

where Δ denotes the true change and π the effect of repeated testing in the members of the control group. For those who are not familiar with CTT, the following analogy may be helpful. Every researcher who applies analysis of variance when comparing group means is familiar with the concept of splitting the total observed variance into the between group variance and the within group variance. This applies to ordinary scores as well as to difference scores. The analogy with analysis of variance reveals when the (conceptual) distribution of possible replications under similar circumstances within a person is regarded as a group of scores, each participant identifying a different group. The group mean is by definition—a basic definition of CTT—the true change (including the non-random test effect) in that person. The variance of the true changes can be regarded as between group variance. The random variance component is the variance of the (conceptually) possible replications within a person. In CCT, these random variance components are assumed to be equal across individuals, and is named error variance of measurement of the difference (thus S^2_{ED}). Considering the definition of an RCI, for the calculation of the standard error only the random error variance within individuals is needed, whereas Collie *et al* also include the between group (= between person) variance. Under the null hypothesis of zero true change and in the absence of varying

effects of testing, the between person variance of the differences is zero, and expressions 3 and 4 prove to be equivalent. Expressions 3 and 4 will also be equal when all the participants in the control sample experience the same practice effect. However, when differential practice effects occur, expression 4 is larger than expression 3.

In addition to this theoretical view, I have shown that the standard error according to Collie *et al* is larger than expression 3, when the initial and final variance observed in the control group differ.[12] The larger the difference between the initial and the final variance and the larger the test-retest reliability, the larger is the difference between the two standard errors.

HOW TO ACT WHEN INITIAL AND FINAL VARIANCE DIFFER

Two subsequent assessments of the same test in a control group without intervention and without the occurrence of test effects ''ideally'' yield equal initial and final group mean, and equal initial and final variance. When these ideal circumstances are met, the distinctive RCIs incorporating expressions 2, 3, or 4 will yield the same outcomes. However, chance has already been mentioned as a cause of differing initial and final variance. Whether a difference between correlated variances can be explained on the basis of chance or not can be tested with a statistic called in from basic text books[6]:

$$t = \frac{(S_X^2 - S_Y^2)\sqrt{N-2}}{2S_X S_Y \sqrt{1 - r_{XY}^2}} \qquad (8)$$

which follows a *t* distribution with df = N−2.

Effects of testing are a second possible cause of differing initial and final variance. Therefore the distinction between the standard errors under review is particularly relevant when measures are used that are susceptible to effects of testing. In sports medicine research, as in neuropsychology research, tests measuring cognitive function are typically used, and effects of testing should then be expected, in particular practice or learning effects. When the regression of the final test score on the initial test score is only influenced by measurement errors, the coefficient of regression b equals the test retest-reliability r_{XY}. When b and r_{XY} differ considerably, effects of testing are also present. Because $b - r_{XY} = r_{XY}(S_Y/(S_X - 1))$, the question of whether of b and r_{XY} differ considerably is linked to the statistical test of the equality of initial and final variance. Therefore, a significant outcome of statistic 8 is regarded as a strong indication that effects of testing are present. RCIs accounting for effects of testing have been introduced in neuropsychology research by Chelune and coworkers[16] and by McSweeny and colleagues.[17] As these two methods have been presented by the same team of researchers, hereafter they will be referred to as method 1 and method 2 respectively. Method 1 retains the standard error of the JT procedure (expression 2), but, in order to account for effects of testing, they modified the numerator—that is, the observed difference score—by subtracting a constant equal to the mean difference score observed in the control group, thus shifting the distribution of observed differences along the score axis into a standardised normal distribution. Method 1 can be regarded as an extension of the classic approach. Method 2 is a regression procedure, because its RCI incorporates (*a*) in the numerator an adjustment to the observed difference score using information on the regression of the observed difference scores on the initial scores, and (*b*) in the denominator the standard error of prediction (of the final score from the initial score).

For the reader, an outline of the choices to be made when assessing the reliability of change may be helpful.

(*a*) When differential effects of testing do not occur, initial and final variance differ only as a consequence of chance, which can be tested with the help of formula 8. In this situation, the regression of the final scores on the initial scores is only influenced by measurement errors, and the regression coefficient b will be close to r_{XY}. Although the use of expression 3 is preferred for conceptual reasons, the outcomes of the various RCIs incorporating expressions 2, 3, or 4 as standard error will be similar.

(*b*) When initial and final variance differ significantly, b and r_{XY} are notably different as well, and in addition to regression to the mean as a consequence of measurement errors, effects of testing are also present. When $b > r_{XY}$, the final variance exceeds the initial variance, and when $b < r_{XY}$, the final variance is smaller than the initial variance. In principle, method 2 should then be used, because the regression method uses the available information more thoroughly.

(*c*) However, because the coefficient of regression of the difference scores on the initial scores is equal to b−1, no useful information can be extracted from this regression when b is approximately 1. In that case, method 2 and method 1 incorporating expression 3 as standard error coincide—even the standard errors coincide—and method 1 incorporating expression 3 can be applied. When the average practice effect in the control group does not significantly deviate from 0, the classic approach can be used—that is, the JT method with expression 3 as standard error.

From the foregoing outline, on the one hand, it can be concluded that method 1 using expression 3 in the denominator is a special case of method 2, which can be viewed as another argument that expression 3 is conceptually preferred. On the other hand, the classic approach using expression 3 is a special case of method 1 using expression 3. Thus, method 1 takes an intermediate position. Many researchers and clinicians will prefer to use this method, when the more elaborate method 2 is not indicated.

COMPARING APPROACHES IN PRACTICE

This discussion of the differences between the standard errors resulting from expressions 2, 3, and 4 will now be illustrated with an example from sport concussion research literature. For this example, outcomes are borrowed from the study by Iverson *et al*.[3] This study involved five composites assessed in a control group where no intervention was conducted, including four scales measuring cognitive function (verbal memory, visual memory, reaction time, procession speed) and a postconcussion symptoms scale. Table 1 shows outcomes already presented by the authors, as well as additionally calculated outcomes (including the standard error according to expression 2 and regression coefficient $b = r_{XY}S_Y/S_X$).

At first glance, the table reveals that the standard error calculated according to expression 4 is indeed always larger than that calculated according to expression 3.

Considering the outcomes for the five tests successively, we see that verbal memory showed no notable increase in the group mean, but did show a notable diminution of the variance, which failed to be significant ($t = 1.66$, $p > 0.10$). As a consequence, the outcome of expression 2 was 8% larger than that of expression 3. The results for visual memory resembled the ''ideal'' outcomes described above and will not be discussed further. Reaction time showed a very small diminution of the group mean but a significant diminution of the variance ($t = 3.94$, $p < 0.001$), resulting in a large difference between the outcomes of expressions 2 and 3, the former being 14% larger, and the largest difference (10%)

Table 1 Descriptive statistics of the healthy control subjects in the Iverson *et al* study[3] (n = 56)

Test	M_x	M_y	M_y-M_x	SD_x	SD_y	r_{xy}	$b_{y.x}$	Standard errors		
								Expr 2	Expr 3	Expr 4
Verbal memory	88.68	88.84	0.16	9.50	8.09	0.70	0.60	7.36	6.83	6.94
Visual memory	78.70	77.48	−1.22	13.39	12.67	0.67	0.63	10.88	10.59	10.61
Reaction time	0.543	0.536	−0.007	0.087	0.063	0.79	0.57	0.056	0.049	0.054
Procession speed	40.54	42.24	1.70	7.64	7.06	0.86	0.79	4.04	3.89	3.93
PSS	5.23	5.79	0.56	6.75	10.07	0.65	0.97	5.65	7.17	7.66

Expr, Expression; PSS, postconcussion symptom scale.

observed between the outcomes of expressions 3 and 4. Procession speed showed a notable increase in the group mean, but no significant change in the variance ($t = 1.14$, $p>0.10$). Iverson *et al*[3] reported significant practice effects as shown by a paired samples *t* test, which led the authors to apply method 1. Although not a cognitive measure, the results for the postconcussion symptoms scale in the control group are also interesting. The mean showed a notable but not significant increase, while the increase in the variance was significant ($t = -2.26$, $p<0.05$), resulting in a 7% larger standard error according to expression 4 and a 21% smaller standard error according to expression 2, as compared with the outcome of expression 3. The value of the test-retest reliability is moderate and the regression coefficient is close to 1, suggesting that the regression to the mean as a consequence of measurement errors is overshadowed by a regression from the mean. The mechanism is not entirely clear, but it possibly evidences greater consciousness of complaints when the questionnaire is completed for a second time, which may be regarded as a specific type of practice effect.

Having applied the classic approach (or in particular method 1 incorporating expression 3 in the case of the processing speed test), Iverson *et al*[3] concluded that ''the percentages of subjects that would be classified as reliably improved or declined was reasonably close to what would be predicted from the theoretical normal distribution''.

However, the authors did not compare the results of applying the classic approach using expression 3 with those of other methods. In this respect, I wish to make several comments. Firstly, Temkin and colleagues,[14 15] who used essentially the same approach as that proposed by Collie and colleagues, reported for the vast majority of the outcome measures included in their research results that did not evidence sufficient variability of the practice effects to yield notably different outcomes for expressions 3 and 4. A striking exception, however, was the trail making test. At the second assessment, this test showed a considerable decrease in the variance, suggesting that this measure was largely affected by variable practice effects. As a consequence, the standard error of the Temkin *et al* approach was 25% larger than that according to expression 3. The Temkin *et al* approach yielded a 90% confidence interval of width 0.83, cutting off only 5% of false positives (whereas 10% was expected on the basis of chance). The use of expression 3 yielded an interval of width 0.67 (number of false positives not reported), whereas the interval really delimiting the most extreme 10% of the participants was of width 0.48, as revealed by a distribution-free approach (table 6 of Temkin *et al*[14]). Thus, as the interval width yielded by expression 3 was far closer to reality, the results were clearly in favour of the classic approach using expression 3.

Secondly, I note that seemingly moderate differences between the sizes of the standard errors and between the numbers of false positives in the control sample may result in considerable differences between the percentages of patients submitted to an intervention, who are assessed as reliably changed.

Thirdly, in the Iverson *et al* study, the regression method was not considered for application. However, for instance, for reaction time, the parameter estimations calculated in the control group possibly indicated the use of this method, which would lead to a standard error 20% smaller than expression 3. The difference between the resulting percentages of patients submitted to an intervention and assessed as reliably changed is probably spectacular.

DISCUSSION

This exposé has led to the conclusion that the standard error proposed by Collie *et al* (expression 4) is always at least equal to or larger than the conceptually preferred expression 3. The distinction between the two standard errors is particularly relevant when the use of the Chelune *et al* method is indicated. The circumstance that the Collie *et al* method is more conservative implies that the probability of an unwarranted designation of reliable change is diminished. This may seem more dependable, but the trade off is a higher risk of not detecting real change. This trade off does not seem to be favourable to sport concussion diagnosis and management. On the one hand, a confidence interval that is too wide prevents a clinician from reliably concluding that the cognitive function of an athlete who has sustained a concussion has deteriorated. On the other hand, a too wide interval prevents a clinician from detecting subtle changes of

What is already known on this topic

In the early 1980s Jacobson and colleagues introduced the reliable change index (RCI) to assess the reliability of within-individual change. An essential element of the RCI formula is the standard error in the denominator, but, since its introduction, confusion has existed about the standard error that should be incorporated. Application of the RCI in sport concussion research and the proposal of a ''new'' formula by Collie and colleagues reflect this confusion.

What this study adds

This article attempts to clarify the confusion. The statistical and psychometric properties of the various RCI formulas are scrutinised and compared, leading to a discussion of which variant might be preferred. A strategy is presented for the procedural choices to be made by practitioners and researchers who wish to establish reliable improvement or deterioration of cognitive function in patients.

cognitive function still present after a period of recovery, enhancing the conclusion that the cognitive function of this athlete has returned to baseline. Thus, if the method proposed by Collie and colleagues is used, there is a greater risk of an unwarranted return of an athlete to the arena after concussion. This risk is admittedly mitigated by the cautious way of using this method advocated by its authors (p 276).[4]

Within this context, it is clear that the suggestion of Collie *et al*,[1] adopted from Hinton-Bayre *et al*,[18] to assess healthy sportsmen twice with alternative test forms at the beginning of the season should be strongly endorsed, especially if methods are pursued that rule out practice effects. Such assessments will provide more accurate establishment of the test-retest reliability and thus of the reliability of change, which is important for methodical soundness and, of course, for the benefit of the sportsmen's health.

As for methods for the assessment of reliable change, other than those compared in the present article, Collie and colleagues[4 5] should be praised for their effort of bringing together current methods from various research areas. However, in my view, their review of these methods did not serve its audience well. The formulas covering the procedures that are in use in the fields with which I am familiar (psychotherapy and neuropsychology research) are highly idiosyncratic. To be honest, I do not understand the denotation of some of them, and I doubt that the readership of this journal does. Furthermore, Collie and colleagues[1 4 5] cited one of my papers,[8] which pointed out serious flaws of two of the methods most extensively cited by Collie *et al*.[4 5] As for the method proposed by Zegers and Hafkenscheid,[19] the latter authors themselves recognised the shortcomings of their proposal, which they no longer endorse.[20] Apart from being based on the concept of Zegers and Hafkenscheid, the proposal of Bruggemans and associates[21] bears a serious mistake and should also not be applied.[8] Most surprisingly to me, one of the methods listed[4 5] was ascribed to me. However, I do not understand this method as denoted by Collie *et al*. I cannot even remember having proposed a new method in the paper cited. On the contrary, I have only attempted to demonstrate that the classic approach presented long ago by McNemar is undeservedly regarded as less dependable for a dependable assessment of change by authors who subsequently presented more sophisticated procedures for a reliable assessment of change.

Competing interests: none declared

REFERENCES

1 **Collie A**, Maruff P, McStephen M, *et al*. Are Reliable Change (RC) calculations appropriate for determining the extent of cognitive change in concussed athletes? *Br J Sports Med* 2003;**37**:370–2.
2 **Hinton-Bayre AD**, Geffen GM, Geffen LB, *et al*. Concussion in contact sports: Reliable Change indices of impairment and recovery. *J Clin Exp Neuropsychol* 1999;**21**:70–86.
3 **Iverson GL**, Lovell MR, Collins MW. Interpreting change on ImPACT following sport concussion. *Clin Neuropsychol* 2003;**17**:460–7.
4 **Collie A**, Maruff P, Makdissi M, *et al*. Statistical procedures for determining the extent of cognitive change following concussion. *Br J Sports Med* 2004;**38**:273–8.
5 **Collie A**, Darby DG, Falletti MG, *et al*. Determining the extent of cognitive change after corony surgery: a review of statistical procedures. *Ann Thorac Surg* 2002;**73**:2005–11.
6 **McNemar Q**. *Psychological statistics*, 3rd ed. New York: Wiley, 1962.
7 **Maassen GH**. Kelley's formula as a basis for the asessment of reliable change. *Psychometrika* 2000;**65**:187–97.
8 **Maassen GH**. Principles of defining Reliable Change Indices. *J Clin Exp Neuropsychol* 2000;**22**:622–32.
9 **Jacobson NS**, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behav Ther* 1984;**15**:336–52.
10 **Jacobson NS**, Revenstorf D. Statistics for assessing the clinical significance of psychotherapy techniques: issues, problems, and new developments. *Behav Assess* 1988;**10**:133–45.
11 **Jacobson NS**, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;**51**:12–19.
12 **Maassen GH**. The standard error in the Jacobson and Truax Reliable Change Index (the classical approach to the assessment of reliable change). *J Int Neuropsychol Soc* 2004;**10**:888–93.
13 **Christensen L**, Mendoza JL. A method of assessing change in a single subject: an alteration of the RC index. *Behav Ther* 1986;**12**:305–8.
14 **Temkin NR**, Heaton RK, Grant I, *et al*. Detecting significant change in neuropsychological test performance: a comparison of four models. *J Int Neuropsychol Soc* 1999;**5**:357–69.
15 **Dikmen SS**, Heaton RK, Grant I, *et al*. Test-retest reliability and practice effects of expanded Halstead-Reitan neuropsychological test battery. *J Int Neuropsychol Soc* 1999;**5**:346–56.
16 **Chelune GJ**, Naugle RI, Lüders H, *et al*. Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology* 1993;**7**:41–52.
17 **McSweeny AJ**, Naugle RI, Chelune GJ, *et al*. ''T Scores for change'': an illustration of a regression approach to depicting change in clinical neuropsychology. *Clin Neuropsychol* 1993;**7**:300–12.
18 **Hinton-Bayre AD**, Geffen GM, McFarland K. Mild head injury and speed of information processing: a prospective study of professional rugby league players. *J Clin Exp Neuropsychol* 1997;**19**:275–89.
19 **Zegers FE**, Hafkenscheid AJPM. The ultimate Reliable Change Index; an alternative to the Hageman & Arrindell approach. *Heymans Bulletin HB-94-1154-EX*. Groningen: Universiteit van Groningen, 1994.
20 **Hafkenscheid AJPM**. De onverbeterlijke neiging tot 'verbetering' van veranderingsscores: terug naar de oorspronkelijke index voor betrouwbare verandering. [The incorrigible tendency to correct change scores: back to the original index for reliable change], *Gedragstherapie* 2002;**35**:317–27.
21 **Bruggemans E**, Van de Vijver FJR, Huysmans HA. Assessment of cognitive deterioration in individual patients following cardiac surgery: correcting for measurement error and practice effects. *J Clin Exp Neuropsychol* 1997;**19**:543–59.

COMMENTARY

The debate over which approach to reliable change is most appropriate to determine significant change in an individual's psychological test scores continues. In a series of papers, Maassen has argued that classical test theory has been unjustifiably ignored by many researchers in the area, and this most recent submission continues this idea. Maassen focuses on what he maintains is an erroneous suggestion of Collie and colleagues,[1] where the latter authors appear to have unknowingly adopted an error term that was earlier proposed by Temkin and colleagues.[2] The interested reader is directed to a series of papers appearing in the *Journal of the International Neuropsychological Society* (volume 10) which consider the issues in more detail (see Maassen[3] and associated dialogues). Although Maassen's discourse is instructive, it is unclear why the regression approach was not presented in more detail. Not only is it proposed by Maassen to be preferable under certain circumstances, but it has also been applied in sports concussion research.[4 5] Using the data of Iverson and colleagues,[6] it can been determined that the error terms for the McSweeny regression approach (standard error of prediction (SEP)) are smaller than any of the other options (expressions 2, 3, and 4) for all tests bar one, the post-concussion scale. This exception is notable as it suggests when the regression approach may be limited—as discussed below. The regression error values are: verbal memory SEP = 5.83, visual memory SEP = 9.51, reaction time SEP = 0.039, processing speed SEP = 3.64, and post-concussion scale SEP = 7.72.

It is my position that the regression approach to reliable change is preferable in most circumstances. I will elaborate on this further, but first to clarify how the RC methods can be practically compared. The RC approaches considered here can all take the general form of RC = (Y−Y′)/SE, where Y is the actual post score, Y′ the predicted post score, and SE a standard error. With a 90% level of confidence, any RC exceeding about 1.645 suggests significant individual change on retest. What varies from one approach to the next is the way in which Y′ and SE are derived. For the original

approach of Jacobson and Truax, Y′ = Y, and SE was originally based on expression 2 (in Maassen's paper), with a recommendation to use expression 3 when posttest values are available, not expression 4. However, as Maassen notes, neuropsychological tests are often subjected to practice/learning effects, and thus this must be taken into account. In response to this problem, Chelune and colleagues[7] calculated Y′ by adding the mean practice effect observed in a control population (or sample) to the pretest score. Maassen again advocates the use of expression 3 over both expressions 2 and 4. Note that the approach of Chelune *et al* adjusts ''practice'' identically for all individuals. Yet, whenever a measure is not perfectly reliable (practically always!), regression to the mean (RTM) will be present. This may be readily determined by obtaining a significant regression of difference scores (Y−X) on pretest scores (X).

The approach of McSweeny and colleagues[8] adjusts for this, such that Y′ is estimated on the basis of the simple Y′ = bX + a equation for a linear prediction. With this approach, more extreme values at pretest are predicted to be closer to the mean on retest, thus each score is adjusted individually on the basis of relative position at pretest. The SEP is used as the McSweeny approach error term. To continue, Maassen suggested that, when practice does not exist and pretest and posttest variances are equal, the approach of Jacobson and Truax with expression 3 as the SE is preferred—for example, in the case of the visual memory data of Iverson *et al*.[6] And, when practice does exist but the variances are still equal, the Chelune method with expression 3 in the denominator is preferred—for example, as seen in processing speed. In both of these situations, as rxy < 1, there will be RTM due to unreliability, which is accounted for by the McSweeny regression approach. So even when pretest and posttest variances are equal ($S_x = S_y$), thus b = r, there is still RTM.

Maassen also suggested that error estimates should converge when variances are equal. However, the SEP is still notably lower than the other SE estimates in both circumstances. In my opinion, this suggests that Y′ should be calculated using the McSweeny approach even when variances are equal.

Maassen goes on to suggest that differential practice exists whenever pretest and posttest variances are not equal ($S_x \neq S_y$, as tested by expression 8). Differential practice can decrease the posttest variance (RTM) or increase the posttest variance (deviation from the mean). These differential practice effects will act in addition to the RTM because of unreliability. Maassen suggests that the McSweeny approach is preferred here, except when b = 1. It is argued that the regression approach is not preferred when b = 1, as it adds nothing to the prediction over the Chelune approach. I agree with this statement, but would interpret its relevance differently. When b approaches 1, as was seen for the post-concussion scale, this suggests that the regression to the mean due to unreliability is ''cancelled out'' by the deviation from the mean, as indicated by the significantly larger posttest variance (differential practice). So adjusting for uniform practice is only warranted when this less common situation occurs—that is, where b = 1 and $S_y > S_x$. This situation can be readily identified when the regression of difference scores (Y−X) on pretest scores (X) is no longer significant (given adequate power), and S_y is significantly larger than S_x—as was observed for the post-concussion scale. So again, whether variances are equal or not, I prefer the McSweeny approach to calculating Y′.

The remaining issue is when expression 3 or SEP is preferred as the SE. When variances are equal, it may be fair to use expression 3. However, when they are not, SEP would be preferable. Moreover, when b = 1, it would seem counterintuitive to combine S_x and S_y in this situation when they have just been demonstrated to be different. Indeed, Iverson and colleagues[6] found that the false positive rate was worst for the post-concussion scale, compared with other measures, with the Chelune approach and expression 3 in the denominator. For these reasons, I prefer the regression based method of McSweeny and colleagues[8] for determining both Y′ and SE in most circumstances.

In summary, because of imperfect reliability, RTM is nearly always significant. This can be tested through the significance of the correlation (or regression slope coefficient) between Y−X and X. Thus, the McSweeny method is preferred for determining the predicted posttest score, even when $S_x = S_y$. Under the circumstance that RTM due to unreliability is cancelled out by deviation from the mean where b = 1 (where $S_y > S_x$, as tested by Maassen's expression 8), the regression method is still preferable, as pooling disparate variances is unwarranted. Needless to say, comparison of false positive rates based on each of these ''theoretical'' approaches would be far more convincing to the clinician. In this regard, I have shown in a limited sample (n = 43) that false positive rates on measures sensitive to the effects of concussion are roughly equivalent for expressions 3 and 4 and the regression approach.[9] In a larger more diverse sample, Temkin and colleagues[2] showed comparable false positive rates for expression 4 and the regression approach—expression 3 was not considered. The data of Barr and McCrea[4] also suggested comparability between the McSweeny and Chelune methods (using expression 2). While we await publication of larger scale studies with various testing outcomes (regarding practice and differential variances), some comfort can be taken that the practical difference between approaches appears to be minor.

A Hinton-Bayre
University of Queensland, School of Medicine, Herston, Qld 4006, Australia; s309339@student.uq.edu.au

REFERENCES

1 **Collie A**, Maruff P, McStephen M, *et al*. Are Reliable Change (RC) calculations appropriate for determining the extent of cognitive change in concussed athletes? *Br J Sports Med* 2003;**37**:370–2.
2 **Temkin NR**, Heaton RK, Grant I, *et al*. Detecting significant change in neuropsychological test performance: a comparison of four models. *J Int Neuropsychol Soc* 1999;**5**:357–69.
3 **Maassen GH**. The standard error in the Jacobson and Truax Reliable Change Index (the classical approach to the assessment of reliable change). *J Int Neuropsychol Soc* 2004;**10**:888–93.
4 **Barr WB**, McCrea M. Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *J Int Neuropsychol Soc* 2001;**7**:693–702.
5 **Erlanger D**, Feldman D, Kutner K, *et al*. Development and validation of a web-based neuropsychological test protocol for sports-related return-to-play decision-making. *Arch Clin Neuropsychol* 2003;**18**:293–316.
6 **Iverson GL**, Lovell MR, Collins MW. Interpreting change on ImPACT following sport concussion. *Clin Neuropsychol* 2003;**17**:460–7.
7 **Chelune GJ**, Naugle RI, Lüders H, *et al*. Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology* 1993;**7**:41–52.
8 **McSweeny AJ**, Naugle RI, Chelune GJ, *et al*. ''T Scores for change'': an illustration of a regression approach to depicting change in clinical neuropsychology. *Clin Neuropsychol* 1993;**7**:300–12.
9 **Hinton-Bayre AD**. Holding out for a reliable change from confusion to a solution: a comment on Maassen's standard error in the Jacobson and Truax reliable change index. *J Int Neuropsychol Soc* 2004;**10**:894–8.