The Two Errors of Using the Within-Subject Standard Deviation (WSD) as the Standard Error of a Reliable Change Index

Gerard H. Maassen*

Utrecht University, The Netherlands

*Corresponding author at: Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands.

E-mail address: g.h.maassen@uu.nl (G.H. Maassen).

Accepted 17 April 2010

Abstract

In this Journal, Lewis and colleagues introduced a new Reliable Change Index (RCI\textsubscript{WSD}), which incorporated the within-subject standard deviation (WSD) of a repeated measurement design as the standard error. In this note, two opposite errors in using WSD this way are demonstrated. First, being the standard error of measurement of only a single assessment makes WSD too small when practice effects are absent. Then, too many individuals will be designated reliably changed. Second, WSD can grow unlimitedly to the extent that differential practice effects occur. This can even make RCI\textsubscript{WSD} unable to detect any reliable change.

Keywords: reliable change index; standard error; within-subject standard deviation

Lewis, Maruff, Silbert, Evered, and Scott (2007) introduced a new Reliable Change Index (RCI\textsubscript{WSD}), using the within-subject standard deviation (WSD) of a repeated measurement design as the standard error. According to these authors, the WSD provides a substantial advantage: As an s.e., it turns out to be notably smaller than the standard errors currently in use. As a consequence, the RCI\textsubscript{WSD} appeared to be much more sensitive in designating individuals as reliably changed. In this note I wish to clarify two major misunderstandings in using WSD. As a start, it is useful to review the RCIs proposed in the literature that are relevant for this discussion. Some of them are currently in use and have also been applied by Lewis and colleagues (2007).

Jacobson, Follette, and Revenstorf (1984) were the first to use the term RCI, proposing a statistic composed as a ratio of an individual’s observed change ($\Delta X$) in the numerator and an s.e. $S_{E_0}$ in the denominator, which they expressed as follows:

$$S_{E_0} = S_X \sqrt{1 - r_{XY}}$$  \hspace{1cm} (1)$$

where $X$ and $Y$, respectively, are the pre- and posttest scores of a certain outcome measure observed in a control group, and $r_{XY}$ is the test–retest reliability. In a letter to the editor of the journal where this RCI was published, Christensen and Mendoza (1986) noted that s.e. (1) was not correct, because it incorporates only the unreliability of one assessment, whereas the assessment of change should involve the unreliability of two measurements. Nevertheless, expression (1) is cited here because it is relevant for the present note, as will be shown later. Complying with the recommendations by Christensen and Mendoza, Jacobson and Truax (1991) proposed the following adjusted standard error:

$$S'_{E_0} = S_{E_0} \sqrt{2} = S_X \sqrt{2(1 - r_{XY})},$$  \hspace{1cm} (2)$$

where $S_{E_0}$ is the s.e. of measurement of the pretest. The reader will see that the correction by Christensen and Mendoza comprehends multiplication of the s.e. by a factor $\sqrt{2}$. The RCI with this s.e. is currently widely used and is cited as RCI\textsubscript{J&T} by Lewis and colleagues (2007). In fact, it is an adaptation of the full formula of the s.e. of measurement of the difference.
score as a function of the variance and reliability coefficient of pre- and posttest, already given by McNemar (1962) in terms of population parameters:

\[ \sigma_{E_0} = \sqrt{\sigma_{E_X}^2 + \sigma_{E_Y}^2} = \sqrt{\sigma_X^2(1 - \rho_{XX}) + \sigma_Y^2(1 - \rho_{YY})}. \]  

S.e.(2) comprehends estimating the population parameters by their values observed in a control sample, assuming that (i) initial and final variance are equal; (ii) the reliability coefficients of pre- and posttest are equal; and (iii) their common value can be estimated by the test–retest reliability. Maassen (2004) has argued that, although the population values of pretest and posttest variance may be equal, pooling the two control sample values provides a better estimate, making it less vulnerable to a deviant pretest variance value in the control sample. Therefore, he proposed

\[ S_{E_0}' = \sqrt{(S_X^2 + S_Y^2)(1 - r_{XY})}. \]  

(4)

Jacobson and colleagues applied their RCIs within the context of psychotherapy research, where it is reasonable to assume that the practice effects do not occur, which is tantamount to assuming equal pre- and posttest variance in the control group. In neuropsychology, however, tests of cognitive function are frequently used and the occurrence of practice effects should be anticipated. Researchers should apply methods of assessing reliable change that account for these effects. Lewis and colleagues (2007) did apply three such RCIs, composed of \( \Delta X - \Delta X_c \) in the numerator and three different standard errors in the denominator. Here, \( \Delta X_c \) is the mean change observed in the control sample, interpreted as the mean practice effect, implying that each individual’s observed change score is corrected for practice effect by subtracting the mean practice effect in the control sample. As a denominator Lewis and colleagues (2007) used:

(a) s.e.(2), leading to an RCI quoted as RCI_{CHELUNE};
(b) the standard deviation of the difference scores in the control sample

\[ S_{\text{diff}} = \sqrt{S_X^2 + S_Y^2 - 2S_XS_Yr_{XY}}. \]  

(5)

The RCI using this s.e. has been quoted as RCI_{ISPFOCD}. S.e.(5) is not consistent with the original definition of assessing a reliable change, since in fact it measures whether an individual change in the experimental group should be considered extreme when compared with the changes established in the control group, regardless of the source of error (random or non-random) (Crawford, Howell, & Garthwaite, 1998, p. 902; Maassen, 2004). The standard errors (4) and (5) are equal if all the practice effects are zero or equal to a constant, because adding a constant to the posttest score then does not change the posttest variance and the correlation between pre- and posttest score. However, when differential practice effects occur RCI_{ISPFOCD} bears the contradiction of incorporating a numerator that is decreased because of the subtraction of an estimation of the practice effect, and a denominator that is increased with the variance of the practice effects (Maassen, 2004). In that case, RCI_{ISPFOCD} is always more conservative than the RCI incorporating s.e.(4).

(c) WSD as denominator, leading to RCI_{WSD}.

Lewis and colleagues (2007) referred to Bland and Altman (1996) who showed how this within-subject standard deviation can be viewed as an s.e. of measurement. To demonstrate the calculation and the applicability of the WSD, I return to Bland and Altman’s note and the data matrix that is central to their text (Bland and Altman, 1996, Table 1). For the present discussion, only the structure of this matrix is relevant, not the content. This matrix shows the scores of 20 persons who were assessed four times, organized in 20 rows corresponding to the 20 individuals, and four columns corresponding to the four consecutive measurements. Bland and Altman argued that the variance of the four scores in each row, calculated as a sample variance (i.e., \( df = 3 \)), can be regarded as estimation of the error variance (i.e., the s.e. of measurement squared) of the test under consideration for the corresponding individual. The s.e. of measurement of a certain outcome measure is usually assumed to be equal across individuals. Thus, the average value of the 20 estimations is regarded as the eventual estimation of this common s.e. squared. This value happens to be the Residual Mean Square (within-subject variance) \( S_w^2 \) in the one-way ANOVA with 20 levels (individuals) and four independent replications per level. Bland and Altman’s argument is correct, provided that the four scores in each row stem from the same distribution, i.e., from a distribution with the same mean, which, according to Classical Test Theory (CTT), is defined as the true score of the corresponding individual. The argument is clearly not correct when the
consecutive scores in a row are affected by effects of testing, which will probably be the case regarding the second, third, and fourth assessment in each row. Consequently, the mean of each row will deviate from the estimate of the true score of a single measurement. Whatever the WSD estimates, it is larger than the estimation of the s.e. of measurement of a single assessment. Unfortunately, Bland and Altman did not emphasize the crucial assumption of zero practice effects, which caused the first misunderstanding in Lewis and colleagues (2007). As noted above, they implemented WSD (\(\frac{S_w}{N}\)) in an RCI that aimed to account for practice effects.

The second misunderstanding in Lewis and colleagues (2007) originates in ignoring the paragraph in Bland and Altman’s text on the difference of two measurements for the same subject. Although Bland and Altman probably were not acquainted with the concept of reliable change, they showed to be aware that the assessment of change should reckon with the unreliability of two measurements. However, as can be derived from the foregoing discussion, by using WSD as s.e., Lewis and colleagues (2007) incorporated only the s.e. of measurement (the unreliability) of one measurement, while the numerator of their RCI contains random error variability of two measurements. Thus, they made the same mistake as Jacobson and colleagues (1984) initially did by proposing s.e.(1), but which mistake was already corrected in the literature about 25 years ago.

Readers who are not familiar with test theory may be puzzled by the theoretical discussion above. They will probably be more convinced by a demonstration of the consequences in practice. From the tables of Lewis and colleagues (2007), it can be ascertained that, for all the outcome variables involved, the WSD is notably smaller than the other standard errors, leading to their claim that RCI\(_{WSD}\) is always more sensitive. Unfortunately, it is impossible to indicate the consequences in the results presented by Lewis and colleagues in terms of percentages of fallible designations, because it is unknown whether the participants really did deteriorate or not. However, it is possible to indicate the consequences of the errors globally and qualitatively, which I will do first.

Although, in principle, WSD can be calculated on the basis of more than two time points, Lewis and colleagues (2007, Table 1) presented the WSD outcomes based on only two measurements for each member of the control group. In that case, WSD can be compared with the other standard errors quoted above. For instance, algebraically it can be easily shown (will not be done here) that the following relationship holds between the squares of s.e.(4) and WSD:

\[
2S^2_w = \frac{N - 1}{N} [S^2_{E_0} + r_{XY}(S_X - S_Y)^2] + \Delta X^2_c. \quad (6)
\]

When the effects of testing do not occur, the mean difference observed in the control sample is expected to be zero, while the initial and final variance are expected to be equal. In that case, the square of s.e.(4), as well as the square of s.e.(2) and (5), is expected to be about twice as large as WSD\(^2\). The WSD is then only about 1/\(\sqrt{2} = 0.7\) times s.e.(2), (4) and (5).

Imagine now the situation where, ceteris paribus (same control group, same outcome variable), all the members of the control group benefit from the same practice effect. [Considering the numerator of the RCIs utilized by Lewis and colleagues (2007), they assumed that this was the case in their samples.] On the one hand, except for RCI\(_{WSD}\), this will leave the RCIs unchanged. Note that the numerator of all the RCIs remains unchanged because adding the same constant to the final score does not change the difference between an individual change and the average change in the control group.

[We have already noted earlier that then the s.e.(2), (4) and (5) also remain unchanged.] On the other hand, this lowers the RCI\(_{WSD}\) outcomes because expression (6) tells us that WSD will grow to the extent that practice effects occur (i.e., to the extent that the difference between initial and final variance and the difference between initial and final mean observed in the control group differ from 0). In principle, when the practice effects are sufficiently large, WSD can even grow to a value that makes RCI\(_{WSD}\) unable to detect the false positives in the control group, and what is worse, reliable changes in the experimental group.

| Table 1. Statistics of the simulated two subsequent assessments in the control sample \((N = 250)\), rendering the outcomes of Tables 2 and 3 |
|-----------------|-------|-------|-------|-------|-------|-------|-------|
| Situation no.   | \(\pi\) | \(M_X\) | \(M_Y\) | \(SD_X\) | \(SD_Y\) | \(r_{XY}\) | \(\text{Standard error}\) |
|                 |       |        |        |        |        |        | Formula (2) | Formula (4) | Formula (5) | WSD  |
| 1               | 0     | 99.43  | 98.75  | 11.25  | 10.91  | .7901    | 7.29       | 7.18       | 7.19       | 4.78  |
| 2               | N(1,0.2) | 100.78 | 101.09 | 11.40  | 6.69   | .5737    | 10.53      | 8.63       | 9.34       | 6.26  |
| 3               | N(5,1) | 100.92 | 104.97 | 11.11  | 7.30   | .6489    | 9.31       | 7.88       | 8.45       | 6.31  |
| 4               | N(25,5) | 100.14 | 125.13 | 12.05  | 9.01   | .5807    | 11.03      | 9.74       | 10.01      | 18.78 |

Note: \(\pi\) = practice effect; the column below \(\pi\) shows the normal distribution of the practice effects.
Numerical Example

We will now illustrate to what extent the general principles discussed above are recognized when analyzing more realistic data. To this end, we created four data files that represent the following situations: (1) negative true change (deterioration) and zero practice effects; (2) negative true change and small practice effects; (3) negative true change and moderate practice effects; (4) negative true change and very large practice effects. The reader will notice that the situations (2) or (3) probably most closely resemble the actual analyses of Lewis and colleagues (2007). Because large-sized control groups is of great importance (Maassen, Bossema, and Brand, 2009), in our experiment we constructed control groups of size $N = 250$, rather than $N = 90$ used by Lewis and colleagues. Each of the four data files comprises an experimental group of size $N = 50$. For every “individual,” an initial score $X$ (pretest) and a final score $Y$ (posttest) are created. Each initial score is the sum of a normally distributed true score ($\mu = 100$, $\sigma = 10$) and a normally distributed measurement error ($\mu = 0$, $\sigma = 5$). The final score $Y$ is the sum of the same true score and a normally distributed measurement error with the same mean and standard deviation. For every individual in the situations (2), (3), and (4), a practice effect is added to the true score component. The practice effect is normally distributed ($\mu = 1$, $\sigma = 0.2$ in situation 2; $\mu = 5$, $\sigma = 1$ in situation 3; $\mu = 25$, $\sigma = 5$ in situation 4) and via a linear regression formula positively linked ($\beta = 0.5$) to the true initial score. This reflects the assumption that individuals with higher initial score more benefit from a practice effect, which is often the case. Finally, for every individual in the experimental group, a true change component is added, which is zero for about half of the individuals, while for the other half a value drawn from the negative half of the normal distribution with $\mu = 0$ and $\sigma = 10$ is added. This leads to negative changes of comparable size (i.e., about the same mean and standard deviation) in the experimental groups of the four situations. (For a formal description of the construction of the data see Maassen et al. 2009.)

Results

Table 1 shows the outcomes of the statistics calculated in the control group that are required for the RCI formulas. Line 1 shows only small differences between s.e.(2), (4), and (5) caused by measurement errors. When practice effects occur (lines 2, 3, and 4), s.e.(2) is always the largest s.e. because the posttest variance then turns out smaller than the pretest variance [see also expression (2)], as a consequence of the positive correlation between pretest score and practice effect. As expected, s.e.(5) is never smaller than s.e.(4). In the situations with zero or small practice effects, WSD is not far from being 0.7 times s.e.(4). In line 4, large practice effects occur, making WSD by far the largest standard error.

Table 2 shows the numbers of type I errors (false positives) in the control group. When using a 90% confidence interval, about 10% false positives are anticipated. In the cases where zero to moderate practice effects occur, RCI$_{WSD}$ delivers much too large numbers of reliable change designations, which is of course a consequence of the relatively small WSD values. In line 4, no false positive is delivered, as a consequence of the large practice effects and the large value of WSD.

In fact, the other three standard errors and corresponding RCIs are only appropriate in situation (1) (no practice effects). In that situation, the three RCI formulas yield exactly the same numbers, being somewhat smaller than the expected 10% designations. In the situations (2), (3), and (4), the three RCIs deliver percentages of reliable changes that more or less differ from 10. The disadvantage of RCI$_{CHELUNE}$ as being dependent on the difference between initial and final variance is particularly revealed in the lines 2 and 3, where the posttest variance is much smaller than the pretest variance and the numbers of reliable change designations are too low.

Table 3 shows the numbers of reliable change designations in the experimental sample, as well as (between parentheses) the numbers of persons that really deteriorated. The results presented in the table mirror the results presented in the previous table: (a) relatively high numbers of reliable change designated by RCI$_{WSD}$ in the situations where zero or small practice effects occur

### Table 2. Members of the control sample ($N = 250$) with zero or positive practice effects classified as reliably changed on the basis of a 90% confidence interval according to various procedures

<table>
<thead>
<tr>
<th>Situation no.</th>
<th>$\pi$</th>
<th>Standard error (2)</th>
<th>Standard error (4)</th>
<th>Standard error (5)</th>
<th>WSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>det</td>
<td>imp</td>
<td>det</td>
<td>imp</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4.0</td>
<td>4.4</td>
<td>4.0</td>
<td>4.4</td>
</tr>
<tr>
<td>2</td>
<td>$N(1,0.2)$</td>
<td>3.6</td>
<td>2.0</td>
<td>6.8</td>
<td>5.6</td>
</tr>
<tr>
<td>3</td>
<td>$N(5,1)$</td>
<td>2.4</td>
<td>3.6</td>
<td>5.2</td>
<td>6.0</td>
</tr>
<tr>
<td>4</td>
<td>$N(25,5)$</td>
<td>4.4</td>
<td>3.6</td>
<td>6.0</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Note: $\pi =$ practice effect; the column below $\pi$ shows the normal distribution of the practice effects. Numbers in Percentages. Abbreviations $det =$ number of subjects classified as reliably deteriorated; $imp =$ number of subjects classified as reliably improved.
and zero designations in the situation with very large practice effects; (b) equal numbers of designations by the other methods when no practice effects occur; (c) smaller numbers of reliable change designations yielded by RCI\textsc{chelune}.

Unfortunately, the table also reveals the shortcomings of the reliable change methodology, being a procedure that aims to unravel several effects from limited information. We know that about half of the “participants” experienced real deterioration. None of the variants is able to detect all these cases. On the other hand, in particular in situation (2), a considerable number of cases have been designated reliably deteriorated while they actually did not change. Moreover, in most of the situations, several cases were designated reliably improved while they actually did not change or even did deteriorate. These findings should be taken as a warning to researchers not to trust unconditionally the outcomes of reliable change assessments by means of the methods here applied.

### Discussion

This commentary responds to an article by Lewis and colleagues (2007) that reports the numbers of patients whose cognitive function did deteriorate after surgery. They have chosen to use two current RCIs: (1) RCI\textsc{j&t}, which expresses the observed change corrected for practice effects relative to the s.e. of measurement (the original concept of reliable change); (2) RCI\textsc{ispod}, which expresses the observed change corrected for practice effects relative to the standard deviation of the changes observed in the control group. In this commentary, we also introduced (3), an RCI advocated by Maassen (2004), which also expresses the observed change corrected for practice effects relative to the measurement error, but which is statistically more appropriate than RCI\textsc{j&t}.

Lewis and colleagues sought “to calculate the RCI by expressing the change in terms of just the random error associated with the assessment”, and they state “Such an estimate of random variance is readily available using the WSD”. What they sought is indeed consistent with the original definition of reliable change assessment. They proposed to use a new RCI, quoted as RCI\textsc{wsd}, accounting for practice effects in the same way as the other three RCIs, but incorporating the within-person standard deviation as standard error.

Our theoretical exposé and the results of our experiment perfectly converge in showing that WSD is inappropriate as an s.e. of an RCI. In the case of two assessments, the use of WSD entails two mistakes, which affect RCI\textsc{wsd} in opposite directions. First, if no practice effects are present, RCI\textsc{j&t}, RCI\textsc{chelune}, and the RCI advocated by Maassen account for the unreliability of two measurements, which is correct when measurement of change is concerned, while WSD is a correct estimate of the s.e. of measurement of only one measurement. When no or small practice effects occur, WSD is much too small as a s.e. of an RCI, which results in delivering much too high numbers of reliable change. Occurrence of small practice effect probably characterizes the data of Lewis and colleagues (2007), leading them to mistakenly boasting that their RCI is substantially more sensitive than the current statistics.

Second, the occurrence of practice effects (equal or differential) enlarges WSD. The growth of WSD is commensurate with the size of the practice effects and WSD may become much larger than the currently accepted standard errors. In principle, this second error may compensate the first error, leading to the results similar to those of the current methods, or even overshadow the first error to an extent that RCI\textsc{wsd} is unable to detect any reliable change.

The RCIs used in this commentary are the same as or similar to the RCI statistics used by Lewis and colleagues (2007). The numerator of these RCIs evolves from the assumption that all the participants benefit from the same practice effect, which is not very realistic. If differential practice effects are anticipated, a paramount procedure is the widely used regression-based approach of McSweeny, Naugle, Chelune, and Lüders (1993). A comprehensive discussion that also presents other options can be found in Maassen et al. (2009).
Conflict of Interest

None declared.

References