LETTER TO THE EDITOR

Neuropsychologists Must Keep Their Eyes on the Reliability of Difference Measures

Bruno Kopp
Cognitive Neurology, University of Technology Carolo-Wilhelmina Braunschweig, Braunschweig, Germany, Department of Neurology, Braunschweig Hospital, Braunschweig, Germany

(Received February 14, 2011; Final Revision February 18, 2011; Accepted February 22, 2011)

Sánchez-Cubillo et al. (2009) conducted a study in which 41 healthy older subjects performed a battery of neuropsychological tests, including the Trail Making Test (TMT), the Digit Symbol subtest (WAIS-III), the Digits Forward and Backward subtests (WAIS-III), a Finger Tapping Test, a Stroop Test, and a task-switching paradigm akin to the Wisconsin Card Sorting Test (cf. Strauss, Sherman, and Spreen, 2006). The results of correlation and regression analyses suggested that TMT A reflects mainly visuo-perceptual abilities, TMT B primarily working memory and task-switching abilities, while the TMTB – A difference score provides a relatively pure indicator of task-switching abilities. The use of the TMTB – A difference score should help clinicians to interpret abnormal performance in terms of a failure of this specific cognitive mechanism.

Unfortunately, there is the danger that the reliability of difference scores will be unacceptably low because the reliability of a difference score is simply a function of the average reliability of its two components and of the correlation between them (Crawford, Sutherland, and Garthwaite, 2008). Under the circumstance of a common standard deviation of the two components used to form the difference, the formula for the reliability of a difference score, r(B – A), is:

\[ r(B - A) = \frac{r(AA) + r(BB) - r(BA)}{2} \]

where \( r(AA) \) and \( r(BB) \) are the reliabilities of the two components, and \( r(BA) \) is the correlation between them (Crawford et al., 2008). Thus, if difference scores compare measures of two related constructs, the correlation between the components will be substantial, and it may eventually approach the reliabilities of the components in its magnitude. Given this situation, the variance of the difference score will predominantly be measurement error variance, simply because the numerator of Equation 1 will approach zero.

The available data point in this direction. Reynolds’ (2002) estimates of internal consistencies and correlations that were obtained in the normative sample of the Comprehensive Trail Making Test (CTMT) are presented in Table 1. Applying Reynolds’ (2002) data to Eq. 1, a reliability of the difference score \( r(B - A) \) = 0.32 is predicted, which is unacceptably low for any clinical purpose. Crawford et al. (2008) obtained even lower estimates for the reliabilities of TMT difference scores from the Delis-Kaplan Executive Functioning System (D-KEFS; Delis, Kaplan, and Kramer, 2001). Specifically, the reliability estimates of the difference score \( D-KEFS-TMT\text{-}TMT\text{ Number–Letter–Switching (TMTB analogue) minus D-KEFS-TMT\text{ Number–Sequencing (TMTA analogue) equaled in three different age groups .10, .06, and .08, respectively (Crawford et al., 2008). Adequate reliability is fundamental whenever the cognitive status of an individual is assessed. The advocated TMTB – A

Table 1. Estimates of CTMT internal consistencies and correlations in the normative sample (N = 1,664; Reynolds, 2002)

<table>
<thead>
<tr>
<th>Trail</th>
<th>Internal consistencies</th>
<th>Correlations (with Trail 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (equals TMTA)</td>
<td>.74</td>
<td>.70</td>
</tr>
<tr>
<td>2 (equals TMTA)</td>
<td>.77</td>
<td>.70</td>
</tr>
<tr>
<td>3 (equals TMTA)</td>
<td>.72</td>
<td>.70</td>
</tr>
<tr>
<td>4*</td>
<td>.70</td>
<td>.61</td>
</tr>
<tr>
<td>5 (equals TMTB)</td>
<td>.70</td>
<td>.59*</td>
</tr>
</tbody>
</table>

Note. TMTA is defined here as a Number-Sequencing task. TMTB is defined as a Number-Letter-Switching task. Trail 1 essentially mimics the TMTA, whereas Trails 2 and 3 are similar to the TMTB but introduce distractor items. The inclusion of distractor items is deemed to be of no relevance at this point. The internal consistency estimates of the three TMTA analogues thus average to .74. Trail 4 of the CTMT does not fit into the TMTA/TMTB dichotomy. In the sample of Sánchez-Cubillo et al. (2009), the correlation between the TMTA and the TMTB scores amounted to .73.
difference score should be considered un-interpretable in those contexts, due to its expectably unacceptable reliability, and, despite its superior construct validity. We deliberately designed the Brunswick Trail Making Test (BTMT) to maximize internal consistency (above the level of .90), by adjusting test length (Kopp, Rösser, and Wessel, 2008). To conclude, neuropsychologists should be very reluctant to use difference scores which compare measures of two related constructs, due to the potential trade-off between gains in validity and losses in reliability. Needless to emphasize that this argument applies to all domains of neuropsychological assessment.

REFERENCES


