Journal of the International Neuropsychological Society (2004), **10**, 902–903.
Copyright © 2004 INS. Published by Cambridge University Press. Printed in the USA.
DOI: 10.1017/S1355617704106127

DIALOGUE

What do Temkin's simulations of reliable change tell us?

GERARD H. MAASSEN
Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, The Netherlands

(Received June 3, 2004; Accepted June 10, 2004)

Due to space limitations I have chosen to confine my reply to the comments by Temkin (this issue, pp. 899–901) that touch most directly the concepts of practice effects and reliable change. Temkin seems to portray my adherence to the classic approach as a private affair. However, Temkin herself (Temkin et al., 1999) reported to utilize the most widely applied procedures of Jacobson and Truax and of Chelune et al., which are based on the classic approach. For unexplained reasons they had substituted a different standard error. The unsatisfactory justification later given in their reply to Hinton-Bayre's (2000) letter revealed the presumably actual reason: unfamiliarity with psychometrics including the classical test theory (CTT). Not surprisingly, Temkin ignores this historical aspect in her comment. Nevertheless, the new *post-hoc* arguments she brings up deserve, of course, a fair evaluation.

In advance, I wish to reply to her remarks on the concept of practice effect. Of course, I know that individual practice effects are usually unknown. Incorporating them in my Expression 1 is only meant to stress that these effects should be accounted for. The principal definition of CTT of a true score (shortly "the score expected under similar circumstances") defines a practice effect as a component of the true change. Of course, neuropsychology researchers want to get rid of this part, and well-known procedures (e.g., Chelune et al., 1993) aim to do so. I am the first to acknowledge that these widely used methods (forcedly) make assumptions which seem questionable in some situations. In fact, I discussed this issue extensively in a recent paper (Maassen, 2003) in which I even suggest improvements.

The core of Temkin's reply are simulation results that claim to reflect real world research. The first simulation involves the sampling of 1,000 cases from a theoretical population of individuals who, apart from practice effects, showed no actual change. From the parameter values provided we can calculate with elementary statistics that the population of difference scores (i.e., practice effects) is normally distributed with mean $.53 - .42 = .09$ and variance $.49{**}2 + .33{**}2 - 2{*}.49{*}.33{*}.83 = .08$. Note that from my Expression 7 it can be derived that this variance is probably larger than the variance of measurement errors. To this population Temkin applies the Chelune procedure with my expression 5* as standard error (i.e., the square root of .08). Establishing RCIs in this way amounts to sampling 1,000 cases from a z distribution. Ten percent of the cases are theoretically expected to show a RCI value (i.e., a z score) outside the interval ±1.645, and that is what she roughly finds. As a consequence of the relatively large variance of Temkin's population, a greater percentage is expected when using the classical (i.e., original Chelune) approach. This does by no means imply that the latter approach is wrong. This simulation only indicates that the percentage of the practice effects that exceed what can be expected on the basis of imperfect measurement (17.5%) is greater than the 10% of the difference scores that are identified as exceptional within this population of relatively large observed differences (which is what Temkin's approach comprehends). Thus, this simulation perfectly illustrates Temkin's misconception of reliable change.

Contrary to the previous simulation, Temkin's second simulation rightly attempts to calibrate measurement errors. However, this attempt shows ignorance of CTT and consequently fails. First, from my paper it can be derived that the initial true scores and the practice effects on the TPT Total Scale are probably substantially correlated. Treating them as independent is presumably the reason that she did not succeed in simulating the actual parameter values. Secondly, from the parameter values specified for the true score variance and the error variance the actual reliability coefficient can be determined (with CTT): $\text{rho} = [.32{**}2]/[.32{**}2 + .01{**}2] = .999$. I emphasize that my Expression 2 is theoretically the correct expression of S_{Ed}, and that my Expression 6 is a way of estimating the correct value when the population parameter values (in particular the reliability coefficient) are unknown. But in this case the actual parameter values are "known," and the unlikely value of

Reprint requests to: Gerard H. Maassen, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: g.maassen@fss.uu.nl.

.999 shows that the parameter values are ill chosen and completely inconsistent with the test–retest reliability estimate of .65. The extremely high value makes every further calculation of Expression 2 unreliable and meaningless. To make such absurd results seem possible, Temkin must have extremely little confidence in the classic approach.

In sum, the two simulations are not valid examples. However, on the other hand, Temkin herself (Temkin et al., 1999, Table 5, under Model 2) presented the corresponding *real* "real life situation" involving the same outcome measure (TPT Total Scale) and the actual parameter values. Applying the Chelune procedure using expression 5* as standard error resulted in a RCI interval that is clearly too wide. Only 4% instead of the expected 10% of the participants were designated reliably changed, which is a reduction of 60%! Unfortunately, she did not present the results using the classic approach. But there is no doubt that they would prove to be better.

REFERENCES

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.

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

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.

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 significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*, 357–369.

https://doi.org/10.1017/S1355617704106127 Published online by Cambridge University Press