Tonio Heidegger
Department of Neurology, Goethe-University Frankfurt,
Schleusenweg 2-16, D-60528 Frankfurt am Main, Germany

Ulf Ziemann
Department of Neurology, Goethe-University Frankfurt,
Schleusenweg 2-16, D-60528 Frankfurt am Main, Germany.
Tel.: +49 69 6301 5739; fax: +49 69 6301 4498
E-mail address: u.ziemann@em.uni-frankfurt.de

Available online 22 March 2011

1388-2457/$36.00 © 2011 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.clinph.2011.02.019

Covariance is the proper measure of test–retest reliability

Test–retest reliability refers to the consistency over time of individual differences on a test (Cook and Beckman, 2006). When comparisons of two separate test sessions are made, test–retest reliability has traditionally been evaluated by means of the Pearson's Product Moment Correlation Coefficient (e.g., Pearson's *r*). Generalizability theory provides a more comprehensive model for assessing reliability (Cronbach et al., 1963). In both cases, test–retest reliability is proportional to the consistent variance in test scores due to individual differences as measured at two or more times.

Song et al. (2011) have recently used an analysis-of-variance (ANOVA) test of group mean differences to evaluate test–retest reliability of the speech-evoked auditory brainstem responses. This analysis provides information of a different sort than is generally understood as reliability in the psychometric literature. A lack of a significant difference between groups provides evidence that there are not systematic changes in test performance over time such as might be due to learning or practice effects. However, this sort of analysis does *not* address whether an individuals' relative ranking in the population is stable over time.

A number of investigators have previously used the correlation coefficient to assess test–retest reliability of electroencephalographic (EEG) features (e.g., Rentzsch et al., 2008; Tusa et al., 1994). Thus, there is ample precedent in the EEG literature for using the correlation coefficient as an index of reliability. While this is not uniformly the case in the EEG literature, we feel that it would be prudent to measure reliability in terms of the consistent variance associated with individual differences.

The variance in test scores can be due to a variety of factors such as measurement error, short-term effects like alertness, and motivation (i.e., the individual's current state) as well as to stable individual differences (i.e., traits). With large samples, measurement error and state effects will cancel in the group mean. However, these are important determinants of error in diagnosis (McFarland and Cacace, 2006). While not captured by an analysis of mean differences, these sources of error in diagnosis are quantified by the correlation coefficient and the Generalizability coefficient.

Analysis of group means over time provides useful information about test performance. However, it does not quantify the consistency of individual differences over time and thus should not be considered a measure of test–retest reliability.

References

Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med 2006;119(165):e7–e16.

Cronbach LJ, Nageswari R, Gleser GC. Theory of generalizability: a liberation of reliability theory. Br J Stat Psychol 1963;16:137–63.

McFarland DJ, Cacace AT. Current controversies in CAPD: from Procrustes' bed to Pandora's box. In: Parthasarathy TK, editor. An introduction to auditory processing disorders in children. New Jersey: Lawrence Erlbaum; 2006. p. 247–63.

Rentzsch J, Jockers-Scherubl MC, Boutros NN, Gallinat J. Test–retest reliability of P50, N100, and P200 auditory sensory gating in healthy subjects. Int J Psychophysiol 2008;67:81–90.

Song JH, Nicol T, Kraus N. Test–retest reliability of the speech-evoked auditory brainstem response. Clin Neurophysiol 2011;122:346–55.

Tusa RJ, Stewart WF, Shechter AL, Simon D, Liberman JN. Longitudinal study of brainstem auditory evoked responses in 87 normal human subjects. Neurology 1994;44:528–32.

Dennis J. McFarland
The Wadsworth Center,
New York State Department of Health,
P.O. Box 509, Empire State Plaza,
Albany, NY 12201,
USA
Tel.: +1 518 473 4680.
E-mail address: mcfarlan@wadsworth.org

Anthony T. Cacace
Department of Communication Sciences & Disorders,
Wayne State University,
Detroit, MI,
USA

Available online 16 March 2011

1388-2457/$36.00 © 2011 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.clinph.2011.01.052

Reply to Test–retest reliability of the speech-evoked ABR is supported by tests of covariance

Drs. McFarland and Cacace have correctly pointed out that the lack of a mean group difference between two sessions does not inevitably equate with test–retest stability at the individual level (McFarland and Cacace, 2011). For example, such a null ANOVA result might occur if half of the subjects demonstrated a large change in one direction (e.g. increase in latency) and half demonstrated an equal but opposite change. They suggest two approaches to better demonstrate intra-individual consistency between two test sessions. We thus have reanalyzed our results using Pearson's product-moment correlations and generalizability coefficients. See Tables 1–3 which show individual-measure Pearson's *r* scores and both individual-measure and clustered generalizability coefficients.

Generalization coefficients $\geqslant 0.8$ and Pearson's $r \geqslant 0.7$ are commonly accepted cut-offs for reliability (Anastasi and Urbina, 1997; Downing, 2004), or for stricter applications such as clinical use, $r \geqslant 0.9$ (Scientific Advisory Committee, 2002). We present a summary of the generalizability coefficient and Pearson's *r* analyses here. In most, but not all cases, the originally-reported response stability was supported by these metrics of covariance.

In summary, generalizability coefficients (both individual and grouped by condition or response cluster) for the 170 ms /da/, revealed that the various RMA measures, the stimulus-to-response correlations (both z' and lag) and the quiet-to-noise response correlations (Z') had $G > 0.9$. Marginally-reliable values (defined as *G*