Bulletin of the Psychonomic Society
1988, 26 (6), 583-586

Intraclass correlations: A two-facet case study and some comments on the concept of reliability

DAVID G. WASTELL and GEOFFREY R. BARKER
University of Manchester, Manchester, England

The decomposition of observed scores into true scores and measurement error provides the classical conceptual framework for the analysis of reliability. The intraclass correlation coefficient (ICC) provides a general reliability coefficient that may be applied in many situations, providing that care is taken to formulate the appropriate ANOVA model. An illustrative analysis is described of a complex design involving two nested sources of error. An appropriate ICC is derived. The value of the information provided by such a reliability study, especially of the variance components, for designing a decision study is discussed. Because ICC is a function of both subject variability and measurement error, difficulties of interpretation can arise. The test theorist's definition of reliability is argued to conflict with intuitive notions of this concept. *Sensitivity coefficient* is proposed as a more appropriate term for the classical reliability coefficient.

The classical approach to the analysis of reliability is based on what Lumbsden (1976) calls the Model T theory of observational data: observed scores are an additive compound of a *true* score plus a random component, the *error* score. Reliability analysis involves the resolution of raw score variance into these two sources, usually using analysis of variance (ANOVA). The intraclass correlation coefficient (ICC), originated by Fisher (1958), expresses reliability as the ratio of true score variance to total variance. There have been numerous technical papers dealing with intraclass correlations (e.g., Bartko, 1966; Shrout & Fleiss, 1979). Lahey, Downey, and Saal (1983), for instance, recently described a method for isolating sources of unreliability in judge × target interactions. ICCs have been used widely as reliability coefficients in education and psychology, as well as in other areas, such as population genetics (Kempthorne, 1957) and medicine (Donner & Eliasziw, 1987; Fleiss, Slakter, Fischman, & Park, 1979).

This paper describes the basic reliability paradigm and illustrates a development of the method to cover a more complex design that involves two sources of error variance. Despite the literature relating to ICCs and the obvious importance of reliability, there is a lack of clarity about what the concept actually means. One encounters, for instance, such statements as "unreliable measurements cannot be expected to relate to other variables" (Shrout & Fleiss, 1979), which are both vague and actually at odds with the manifest success of psychology at discovering lawful relationships despite endemic and often overwhelming error variance. We believe that there are problems with the test theorist's definition of reliability; we discuss these issues in the final section.

David G. Wastell is with the Faculty of Medicine Computational Group. Geoffrey R. Barker is with the Department of Oral and Maxillo-Facial Surgery. Address reprint requests to D. G. Wastell, Computational Group, Medical School, University of Manchester, Oxford Road, Manchester M13 9PT, England.

A Case Study: Mean Peak Frequency of the Electromyogram

The electromyogram (EMG) is a familiar psychophysiological tool with a long tradition of use in psychology, primarily as an index of muscle tension (Goldstein, 1972). EMG is customarily analyzed in the time domain, often as a chronometric index to measure the onset, intensity, and time course of motor activity (Grunewald-Zuberbier, Grunewald, Runge, Netz, & Homberg, 1981). Useful frequency information is also available in the power spectrum of the EMG, which breaks down signal power into a range of frequency components. In particular, the mean power frequency (*MPF*) of the EMG, which measures the middle of the spectrum, has emerged as a useful parameter in a number of contexts. It is established that *MPF* shifts from high to low frequencies during fatigue (Mills, 1982; Naeje & Zorn, 1981), and there is growing clinical interest in *MPF* as a diagnostic aid in certain facial pain syndromes, where etiology is believed to involve psychogenic factors and muscle hyperactivity (Barker, 1985). Despite some work (Viitasalo & Komi, 1975), doubts remain in the clinical literature about the reliability of *MPF* (Barker, 1985).

This paper reports a two-facet reliability study (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to examine the reliability of *MPF* within and between sessions. Between-sessions stability is of critical importance for electrophysiological indices. Goldstein (1972) observed, referring specifically to the EMG, that "without some constancy from one experimental session to the next, any system of measurement becomes meaningless" (p. 339). We thus took a series of electromyographic recordings on a group of 10 normal subjects, sampled from the staff of the University Dental Hospital, Manchester. The sub-

jects attended three recording sessions, and four measurements of *MPF* were made on each visit using standard procedures (see Barker & Wastell, 1988). Assessments were made of the two main jaw-elevating muscles, the masseter and the anterior temporalis, giving MPF_M and MPF_T.

The Method of Intraclass Correlations

Reliability is normally considered in the context of a paradigm in which a number of supposedly comparable measurements are made on a number of subjects (Winer, 1971). Often the comparable measurements involve two or more judges or tests. In the simplest case, letting X_{ij} denote the jth observation on the ith individual, we have the usual one-way random effects ANOVA model:

$$X_{ij} = \mu + S_i + E_{j(i)} \quad i = 1 \ldots n \quad j = 1 \ldots p, \quad (1)$$

where S_i is the true score for the ith subject, $E_{j(i)}$ is the error associated with observation (i,j), and the notation $j(i)$ means j is nested under levels of i. The calculation of ICC involves estimating variance components from the between-subjects (*MSB*) and within-subjects (*MSW*) mean squares. *MSB* is a compound of true individual differences and error, that is, $p\sigma_S^2 + \sigma_E^2$; *MSW* reflects only σ_E^2. ICC is defined as $\sigma_S^2/(\sigma_E^2 + \sigma_S^2)$ and is readily calculated by solving simple simultaneous equations to determine σ_S^2 and σ_E^2 (the F ratio *MSB*/*MSW* provides a test of statistical significance).

ICC thus measures the ratio of true score variance to total variance. When true score variance predominates over error variance, ICC will be close to 1, and vice versa. ICC can also be interpreted as a correlation coefficient, as indexing the degree to which the comparable measurements arrange the subjects to be judged in the same relative positions vis-à-vis each other. For two sets of comparable measurements (e.g., two judges), ICC is equivalent to the Pearson product-moment correlation coefficient (Bartko, 1966).

Development of a Reliability Model for the *EMG* Study

More complex models and procedures apply where the judges or tests factor is crossed with subjects. Bartko (1966) considered all three models that involve one source of error variance and derived formulations of ICC that preserve its meaning as a correlation coefficient. Of course, none of these models applies to the present case of two differentiable error factors, between and within sessions.

Let us thus begin by developing an appropriate ANOVA model. In particular, what sort of factors are the between-sessions (B) and the within-sessions (W) factors: are they fixed or random factors, nested or crossed? The question of order effects arises in any situation involving repeated testing. Cronbach et al. (1972) discuss this issue (p. 176) in their treatment of the Porch test for aphasic patients. A preliminary analysis treating both B and W as crossed factors revealed no significant order effects. Thus both were treated as random factors, nested within subjects and sessions, respectively.

This gave the following hierarchical ANOVA model:

$$X_{ijk} = \mu + S_i + B_{j(i)} + W_{k(ij)}$$
$$i = 1 \ldots n, \; j = 1 \ldots p, \; k = 1 \ldots q, \quad (2)$$

with $n=10$, $p=3$, and $k=4$. S_i reflects the true score for subject i, $B_{j(i)}$ the session effect for his/her jth session, and $W_{k(ij)}$ the error for the kth observation in that session. Because there are no interactions in hierarchical ANOVA, the breakdown of variance is very straightforward; the variance components of the mean squares are given in Table 1. From the definition of ICC as the ratio of true score variance to the variance of observed scores, we have, for the present design, ICC $= \sigma_S^2/(\sigma_S^2 + \sigma_B^2 + \sigma_W^2)$. It is straightforward, by following the general approach laid out by Bartko (1966), to show that this formulation can be legitimately interpreted as a correlation coefficient (see Appendix). *MSS*/*MSB* provides an appropriate test of significance.

The ANOVA summary table for MPF_M is shown in Table 2. The following estimates of σ_S, σ_B, and σ_W were derived: 6 Hz, 7.1 Hz, 5.4 Hz. ICC was thus 0.31 ($p < .05$). The corresponding variance components for MPF_T were 9.3, 7.1, and 5.4 Hz; thus ICC was 0.52 ($p < .01$).

Discussion of Results: Planning the *D*-Study

At the heart of Cronbach et al.'s (1972) seminal work on reliability is the fundamental distinction between a generalizability study (G-study) and a decision study (D-study). The role of a G-study, like the present study, is to provide as much information as possible to aid the planning of a D-study. The emphasis of the G-study, therefore, should be on measuring as many facets of variation as may be relevant. Of the various facets examined

Table 1
Random Effects, Hierarchical Analysis of Variance: Within-Sessions Variation (W) Nested in Between-Sessions Variation (B) Nested Within Subjects (S)

Source	*df*	*MS*	*E*(*MS*)
Subjects	$n-1$	*MSS*	$\sigma_W^2 + q\sigma_B^2 + pq\sigma_S^2$
Between Sessions	$n(p-1)$	*MSB*	$\sigma_W^2 + q\sigma_B^2$
Within Sessions	$np(q-1)$	*MSW*	σ_W^2

Table 2
Analysis of Variance for MPF_M: Within-Sessions Variation Nested in Between-Sessions Variation Nested Within Subjects

Source	*SS*	*df*	*MS*	*F*
Between Subjects	6,064	9	673.9	2.79 $p < .05$
Between Sessions	4,832	20	241.6	
Within Sessions	2,334	90	25.9	
Total	13,230	119		

here, several (i.e., the order effects) appeared to be of little significance, leaving a final model involving only the two random effects. What information do these two variance components convey in relation to the design of a D-study?

Consider the within-sessions error first. Our variance estimate for this component was relatively low, but it is a nuisance worth eliminating for the sake of a few extra observations. Let us make our first design decision to record four *MPF* values per recording session and to take the average. The overall error variance (σ_E^2) in the contemplated D-study is thus $\sigma_B^2+\sigma_W^2/4$, and we may obtain an estimate of 7.6 Hz for σ_E from the variance components estimated in our G-study. We may also estimate the reliability of the D-study, in this case ICC = 0.38. Between-sessions variance puts a low ceiling on further improvements that could be wrought by increasing the number of *MPF* measurements per session. Let us therefore contemplate the gains of a design involving more than one session. If subjects attended two sessions (for say eight readings), the estimated error would be 5.4 Hz and reliability 56%. Having subjects attend more than one session poses obvious logistical and economic difficulties, but the projections from our G-study allow these costs to be rationally weighed against the statistical benefits.

Intraclass Correlations: Reliability or Sensitivity?

It is axiomatic in developing psychological, educational, or clinical tests—indeed, tests of any sort—that measurement be reliable. Central to the intuitive concept of reliability is the elementary notion that we may depend upon our test to give reproducible results, that is, that repeated applications under constant conditions should yield comparable results. Recall, for example, Goldstein's (1972) desideratum that measurement should show "constancy from one experimental session to the next" (p. 339). Nowhere does the test theorist's concept of true scores appear in this intuitive notion of reliability. We believe that there are conceptual problems with the definition of reliability as a ratio involving true score variance, symptoms of which break out when we come to interpret actual values of ICC.

Let us consider our results. The value of ICC for MPF_M looks low; it indicates over twice as much error variance as true score variance. The problem is that because ICC is a ratio, there are two explanations for low reliability: either measurement error is high or true score variance is low (Lahey et al., 1983). Consider now our finding that MPF_T is more reliable than MPF_M. This is not because the error components are lower, but because there is greater variation among people. These examples illustrate a tension between the intuitive concept of reliability-as-repeatability, which corresponds directly to the simple concept of error variance, and the test theorist's definition, which incorporates the abstract (and ontologically problematic) notion of a true score. Crudely speaking, one feels uneasy about a measure of reliability that is a function of between-subject variability, especially when this same variability will be a source of classification error when we come to use the test in selection, appraisal, or diagnosis.

One is reluctant to advocate new terminology, especially in an area in which some schism has already occurred (Cronbach et al., 1972), but we believe that many of the foregoing difficulties can be resolved by a simple change of name. ICC is a ratio: it is a ratio involving the magnitude of some quantity to be measured, usually differences among people, and sources of variation that impede such discrimination. In short, ICC is a sort of signal-to-noise ratio; it measures the *sensitivity* of our measuring instrument relative to the signal to be detected and the prevailing noise. Let us then dub the intraclass correlation a *coefficient of sensitivity*. By this simple expediency, we have a more appropriate definition and one that avoids conflict with intuitive notions of reliability. Cronbach et al. (1972) also allude to the problems posed by ratios and argue that the emphasis in G-studies should be on the reporting of variance components rather than reliability coefficients.

REFERENCES

BARKER, G. R. (1985). *Frequency domain studies of the masseter and anterior temporalis electromyograms*. Unpublished master's thesis, Manchester University Library.

BARKER, G. R., & WASTELL, D. G. (1988). The effect of fatigue on the silent period of the masseter electrogram. *Journal of Dentistry*, **16**, 71-75.

BARTKO, J. J. (1966). The intra-class correlation coefficient as a measure of reliability. *Psychological Reports*, **19**, 3-11.

CRONBACH, L. J., GLESER, G., NANDA, H., & RAJARATNAM, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DONNER, A., & ELIASZIW, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine*, **6**, 441-448.

FISHER, R. A. (1958). *Statistical methods for research workers*. New York: Hafner.

FLEISS, J. L., SLAKTER, M. J., FISCHMAN, S. L., & PARK, M. H. (1979). Inter-examiner reliability in caries trials. *Journal of Dental Research*, **58**, 604-609.

GOLDSTEIN, I. B. (1972). Electromyography: A measure of skeletal muscle response. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 329-365). New York: Holt, Rinehart & Winston.

GRUNEWALD-ZUBERBIER, E., GRUNEWALD, G., RUNGE, H., NETZ, J., & HOMBERG, V. (1981). Cerebral potentials during skilled slow positioning movements. *Biological Psychology*, **13**, 71-87.

KEMPTHORNE, O. (1957). *An introduction to genetic statistics*. New York: Wiley.

LAHEY, M. A., DOWNEY, R. G., & SAAL, F. E. (1983). Intraclass correlations: There's more than meets the eye. *Psychological Bulletin*, **93**, 586-595.

LUMBSDEN, J. (1976). Test theory. *Annual Review of Psychology*, **27**, 251-280.

MILLS, K. R. (1982). Power spectral analysis of electromyogram and compound muscle action potential during fatigue recovery. *Journal of Physiology*, **326**, 401-409.

NAEJE, M., & ZORN, H. (1981). Changes in the power spectrum of the

surface electromyogram of the human jaw muscles during fatigue. *Archives of Oral Biology*, **26**, 409-412.

SHROUT, P. E., & FLEISS, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420-428.

VIITASALO, J. H. T., & KOMI, P. V. (1975). Signal characteristics of EMG with special reference to reproducibility of measurements. *Acta Physiologica Scandanavica*, **93**, 531-539.

WINER, B. J. (1971). *Statistical principles in experimental design*. Tokyo: McGraw-Hill.

APPENDIX

To show that $\sigma_S^2/(\sigma_S^2+\sigma_B^2+\sigma_W^2)$ has the structure of a correlation coefficient, begin by considering any two observations X_{iab} and $X_{ia\beta}$ for a given person i. Using Equation 2 and taking expectations, we may prove

$$\begin{aligned}\mathrm{Cov}(X_{iab},X_{ia\beta}) &= E\{[X_{iab}-E(X_{iab})][X_{ia\beta}-E(X_{ia\beta})]\}\\ &= E\{[S_i+B_{a(i)}+W_{b(ia)}][S_i+B_{\alpha(i)}+W_{\beta(i\alpha)}]\}\\ &= E\{S_i^2\} = \sigma_S^2\end{aligned}$$

since all other terms in the expansion have expectation 0.

The total variance of an observation is $\sigma_S^2+\sigma_B^2+\sigma_W^2$. Thus the correlation between X_{iab} and $X_{ia\beta}$ is given by

$$\mathrm{Cov}(X_{iab},X_{ia\beta})/\{\sqrt{\mathrm{Var}(X_{iab})}\sqrt{\mathrm{Var}(X_{ia\beta})}\} = \sigma_S^2/(\sigma_S^2+\sigma_B^2+\sigma_W^2)$$

(Manuscript received January 25, 1988.)