RECENT PROPOSALS TO ESTIMATE TRANSIENT ERROR WITHIN THE CLASSICAL TEST THEORY TRADITION

G.K. HUYSAMEN
gerthuysamen@mweb.co.za
Gordon Institute of Business Science,
University of Pretoria,
and
Department of Psychology,
University of the Free State,

ABSTRACT

Reliability is conceptually defined in terms of consistency across test occasions but coefficient alpha, the most popular reliability estimation method, precludes the examination of such consistency. Three recent proposals to estimate transient error separately within a classical test theory tradition, and the results that they have yielded are reviewed. The merits of these proposals are compared with those of generalisability theory which differentiates between different sources of error variation. Although the procedures reviewed cannot match the advantages of generalisability theory, they may be sufficient in many applications.

Key words
Transient error, test-retest alpha, generalisability theory

The origin of the concept of measurement reliability is usually traced back to the contributions of Charles Spearman and Edward L. Thorndike in the first decade of the 20th century (cf. Crocker & Algina, 1986; Stanley, 1971). In the intervening century, this topic has become increasingly complex as evidenced by R. L. Thorndike (1951), Stanley (1971) and Feldt and Brennan (1989) in the successive editions of the encyclopedic *Educational Measurement*. At the same time, Li, Rosenthal and Rubin (1996) consider reliability to be one of the fundamental notions common to all fields of psychology so that "an understanding of (it) helps define us as psychologists" (p. 98). A clear understanding of this complex topic is important not only for applied psychologists who wish to justify their conclusions made on the basis of tests or other measurement procedures; it is equally important for all researchers who use such instruments to assess variables of interest to their research. Reliability derives its importance from the statistical limit it imposes on validity (Schmidt & Hunter, 1996). To the extent that a variable is not measured perfectly reliably, that is, to the extent that measurement error is present, its observed correlation with any other variable is attenuated. As a consequence, the magnitude of effect size estimates is reduced and the probability of obtaining statistically significant results is diminished. Wilkinson and the APA Task Force on Statistical Significance (1999) deemed the effect of reliability on the magnitude of effect size estimates so important that they made an assessment of reliability mandatory for the interpretation of such estimates.

If one considers the manner in which reliability information is presented in most research reports, one may be forgiven for getting the impression that reliability is a fixed, immutable property that inheres in a particular measuring instrument, and that it may be estimated equally acceptably by different methods. Instead, there are several sources of measurement error (e.g., test items, test occasions, their interactions with test participants, random response) which may differ in their respective attenuating effects on reliability. This multifaceted nature of measurement error is certainly not new. Six and a half decades ago, Hoyt (1941) and Jackson and Ferguson (1941) introduced the estimation of separate error variances within an analysis-of-variance framework. A few decades later, Gleser, Cronbach and Rajaratnam (1965) and Cronbach, Gleser, Nanda and Rajaratnam (1972) further developed this approach into a fully-fledged theory, known as generalisability theory. However, the continued reliance on methods that estimate a single, undifferentiated error variance may reinforce the notion that the error arising from different methods is overlapping or substitutive rather than cumulative. In the process the true magnitude of some sources of error variation, particularly that due to transient error, may go undetected.

Recently, three procedures have been developed within the classical test theory tradition to remedy this situation. The general purpose of this article is to revisit the multifaceted nature of measurement error and to review these approaches, to report the empirical evidence on the potency of transient error as revealed by these methods, and to compare these new approaches with generalisability methodology. However, first a brief review of reliability theory and estimation is in order to provide a framework for the remainder of the article.

DEFINITION AND ESTIMATION OF RELIABILITY

Reliability may be defined either statistically or in nontechnical terms. Whereas the latter is designed to convey an intuitive, conceptual understanding of this concept, the statistical definition makes it possible to quantify it. First the statistical definition will be presented and later two definitions that are formulated in laypersons' terms will be quoted.

In classical test theory an individual's observed test score is decomposed into only two undifferentiated sources of variance:

$$\text{Observed score} = \text{True score} + \text{Error Score}. \qquad (1)$$

The true score in classical test theory is defined as the mean of all the scores a person would have obtained if he or she has taken the test an infinite number of times, each time attempting the test as if for the first time (i.e., in the absence of any transfer effects such as practice, memory or fatigue). This notion of a true score as the constant, systematic component of an individual's observed scores over such independent repetitions, called replications, should be distinguished from what has been referred to as the platonic notion. The latter concept refers to a person's actual standing on the attribute in question, as supposedly known by some omniscient being.

For any particular individual the observed score may be expected to fluctuate around his or her (constant) true score, sometimes exceeding it and sometimes falling below it. The resulting deviation of the observed score from the true score is known as the error score. Just as the present definition of a true score should be understood as a statistical entity, so error scores should not be interpreted as mistakes but merely as residual scores, that is, as the deviations of observed scores from true scores. Suppose the practice examples of a maximal-performance test, which all test participants typically answer correctly, are inadvertently counted as part of the total test score. As this (constant) marking error would be incorporated into the true score, it would mean that all participants' observed scores and true scores are increased by a constant (equal to the marks earned by the practice examples), but that the error scores would remain the same.

Now, if a population of individuals were to take a test repeatedly an infinite number of times without there being any transfer from one test application to the next, each of the components in Equation (1) may be expected to show some variance across individuals. Although the true score is a constant for any particular individual, it has different values for different individuals and so generates a (true-score) variance. Under certain assumptions (cf. Lord & Novick, 1968) the variance of the observed scores will then be equal to the sum of the variances associated with each of the two components on the right-hand side of Equation (1):

$$\sigma^2_{\text{observed}} = \sigma^2_{\text{true}} + \sigma^2_{\text{error}}. \qquad (2)$$

In terms of this exposition, then, reliability is defined statistically as the ratio of true-score variance to observed-score variance, which may be interpreted as the proportion of observed-score variance that is attributable to true-score variance,

$$\rho_{\text{rel}} = \sigma^2_{\text{true}}/\sigma^2_{\text{observed}}. \qquad (3)$$

If one solves for σ^2_{true} in Equation (2) and substitutes the result into Equation (3), we obtain:

$$\rho_{\text{rel}} = [1 - (\sigma^2_{\text{error}})/(\sigma^2_{\text{observed}})]. \qquad (4)$$

Equations (3) and (4) define reliability in terms of quantities that are not directly observable. The concept of parallel tests makes it possible to estimate reliability empirically. Any particular individual's true score (i.e., his or her mean over replications) on two strictly parallel tests is defined to be equal. This means that strictly parallel test forms or, simply, parallel tests, are equally difficult or equally attractive (in typical-performance tests) for a population of individuals. Every individual in a population of persons also has equal error-score variances and, as a result, equal observed-score variances over replications of such tests (as $\sigma^2_{\text{true}} = 0$ for any given individual), implying that whatever these tests measure, they do it equally well. The error scores on one parallel test are uncorrelated (over participants) with those on another parallel test and uncorrelated with the true scores on another parallel test. In terms of the foregoing, reliability may be defined statistically as the correlation between strictly parallel tests, *X* and *X'*.

$$\rho_{\text{rel}} = \rho_{XX'}.$$

Against this statistical background, let us consider two nontechnical definitions of reliability found in the introductory textbook of Anastasi and Urbina (1997, p.8), namely, that

> Test reliability is the consistency of scores obtained by the same persons when retested with the identical test or with an equivalent form of the test,

and that of Crocker and Algina (1986, p. 105), stating that,

> ...(R)eliability is the degree to which individuals' deviation scores, or standard scores, remain relatively consistent over repeated administration of the same test or alternate test forms.

In their treatment of reliability estimation, textbooks would typically refer to the test-retest method, the parallel-forms method and various internal-consistency methods, which include the split-half methods and coefficient alpha (Cronbach, 1951). In the test-retest method, the same test is administered on different occasions and the correlation between the two data sets so obtained is known as the *coefficient of stability*. To the extent that there is transfer from the first to the second application of the same test, or to the extent that participants (incorrectly) interpret the second administration as a test of their consistency in responding, the two applications of the same test do not meet the independence assumption referred to above. As a result, the interval between the two test occasions should be long enough to eliminate possible transfer effects but not too long so as to avoid permanent changes from occurring in the attribute being measured. However, as the content of different parallel forms is not the same, such tests may be applied in close succession as no extended interval is necessary to eliminate memory or false consistency effects. The correlation between the two data sets obtained by administering parallel forms on the same occasion is known as the *coefficient of equivalence*. The correlation between parallel test forms administered with a time interval in between is referred to as the *coefficient of equivalence and stability*.

Split-half coefficients, derived under the assumption that the test halves are parallel, or essentially tau-equivalent, and coefficient alpha, which assumes essentially tau-equivalent items, are coefficients of internal consistency. Unlike parallel components (items or subtests), essentially tau-equivalent components may show different true scores for any particular individual, but for all individuals these true scores differ by the same additive constant: if this constant is, say, +0,85, for every individual the true score on the one component will be 0,85 points higher than that on the other. These assumptions make it possible to estimate the reliability of a full-length test in terms of parts (halves or items) of the test. In the case of coefficient alpha (or the Rulon-Flanagan split-half methods), the obtained coefficient is also an estimate of the correlation between the existing test and another so-called randomly parallel test. Randomly parallel tests are comprised of random samples of comparable items (Cronbach, 1951).

Hogan, Benjamin and Brezinski (2000) studied the frequency with which these various types of reliability methods had occurred among all the reported reliability estimates in the American Psychological Association's Directory of unpublished experimental mental measures. They found that coefficient alpha was used in two-thirds of all the cases, that test-retest reliability was determined in slightly less than one-fifth, with all of the other methods featuring in less than 5 per cent of the cases. The reason for the popularity of coefficient alpha is obvious: it requires that only one form of the test, typically the only form available, be administered only once. Moreover, its use exempts one from making the complicated decision on how to split the test into equivalent halves as is required in the case of the split-half methods.

The test-retest method requires considerably more effort than the internal-consistency methods as it requires that the same test be administered twice. Apart from the extra time taken up by such repetition, there is the difficulty of collecting all of the test participants from the first administration for the purposes of the second administration. The Hogan et al. review found no reports in which the parallel-forms method was used. This is not surprising considering that this method exacts an even greater input in terms of time and especially finances as it requires that another test be constructed so that the two tests meet the statistical requirements of strict parallelism. Strictly parallel test forms are sometimes required to double-check the unconvincing performance of participants on an earlier application, or to gauge, by means of a posttest, the effectiveness of an intervention that was designed to induce change (in the attribute being assessed). However, it may be considered to be unnecessarily restrictive to construct a strictly parallel test solely for the purposes of investigating reliability.

As the estimation of the coefficient of equivalence and stability additionally requires that the two test forms be administered on separate occasions, it is the most labour-intensive of all the reliability estimation methods.

ESTIMATING UNDIFFERENTIATED ERROR THAT TYPICALLY EXCLUDES TRANSIENT ERROR

The *or* in the two nontechnical definitions above may be interpreted to mean that reliability is an immutable property that may be estimated by any of the methods mentioned above and that the results of these methods are interchangeable. This would be the case only if, in terms of Equations (1), (2) and (4), the measurement error estimated by one method were the same as that estimated by any other. However, a moment's reflection would reveal that different estimation methods are susceptible to different kinds of measurement error. Apart from random response error, the major sources of measurement error in psychological testing are transient error and specific-factor error (Schmidt, Le & Ilies, 2003). (R.L.Thorndike, 1951, provides a more exhaustive catalogue of potential sources of measurement error.)

Random response error is due to momentary fluctuations in test participants' responses that cause the same individual to provide different answers to even comparable items on a particular application of the same test. Whereas random response error manifests itself across items administered on the same test occasion, transient error refers to sources that affect participants' test scores in different ways on different test occasions. This kind of error may be due to the mood or mental agility of the respondent at the occasion he or she completes the test. On one occasion the test participant may be in an upbeat mood and readily tackle and solve questions that he or she may miss on other occasions when, say, his or her blood sugar level is below normal. Moreover, any particular individual's mood or readiness status will affect his or her responses to all items on a particular test occasion causing him or her to respond more similarly to items within an occasion than across occasions. In analysis-of-variance terminology this means that a participant-occasion interaction is present.

Specific-factor error sources involve those that are specific to the content of a particular test form (but not to that of another). To the extent that the sample of content reflected in a particular test is more familiar to some participants than to others of the same standing on the attribute being measured, specific-factor error variance is present. (The term specific factor derives from factor analysis where a common factor is defined by several variables whereas a specific factor is characterised exclusively by a single variable.) Due to a chance familiarity with, for example, the tools appearing in the items in one mechanical reasoning test, one participant may obtain a higher score on that test than on another test that equally effectively could have been used in the place of the present one. Another person, again, may disproportionately benefit from the inclusion of content in the latter rather than the former test so that a participant-test form interaction is present.

So, a more representative picture of what is at stake may be obtained by decomposing the undifferentiated error score component in Equation (1) as follows:

> Observed score = True score + Transient error score + Specific-factor error score + Random response error score.

As a result, the undifferentiated error-score variance, σ^2_{error} , in Equation (2) should be replaced by separate variance terms for each of transient, specific-factor and random response error (providing, of course, that the data collection design involved allows for their separate estimation), and Equation (4) should be rewritten as follows:

$$\rho_{\text{rel}} = [1 - (\sigma^2_{\text{transient error}} + \sigma^2_{\text{specific-factor error}} + \sigma^2_{\text{random-response error}})/(\sigma^2_{\text{observed}})] \qquad (5)$$

The inability to estimate any kind of error means that the corresponding component in the numerator of the ratio on the right-hand side of Equation (5) is not estimated. As a result, the right-hand side of the entire equation, that is, estimated reliability, is increased. Stated in terms of Equation (3), the variance of the error component that is not assessed by a particular method is incorporated into the true-score variance, resulting in an overestimate of reliability yielded by that method.

Now, in the test-retest method the same test is administered on different occasions so that transient error complements the error-score variance in Equation (5) and so reduces the coefficient of stability. But because this method involves the same test form, it is not susceptible to specific-factor error, so that variance due to the latter is missing from the error variance in Equation (5) but is incorporated into the true-score variance in Equation (3) and, hence, increases reliability in terms of both equations. (Moreover, in the test-retest method memory effects will spuriously raise the obtained correlation whereas permanent changes of different degrees in different participants will spuriously lower it.) Because parallel test forms contain different content, the coefficient of equivalence, in common with the coefficients of internal consistency, including coefficient alpha, is susceptible to specific-factor error and consequently is attenuated by it. If the two tests forms are administered under the same temporary factors, this coefficient is unaffected by transient error as it assigns transient error to true-score variance, resulting in a higher value for Equation (3). However, the coefficient of equivalence and stability is depressed by the effects of both transient and specific-factor error. Feldt and Brennan (1989) give the exact formula for the reliability coefficient for each of these situations and these formulae show the sources of variation they respectively treat as true-score and error variation.

From the above it follows that the reliability estimates resulting from different estimation methods are reduced by different sources of measurement error. Consequently, an estimate that doesn't take account of all of them, results in an overestimate. Becker (2000) uses the terms *complete reliability* and *partial reliability* to distinguish between estimates that are susceptible to both specific-factor error and transient error, on the one hand, and estimates that are affected by specific-factor error only (i.e., apart from random response error), on the other hand. As neither the coefficient of stability, nor the coefficient of equivalence, or any of the internal-consistency coefficients, are susceptible to both kinds of measurement error, all of them fail to reflect complete reliability. As the coefficient of stability and equivalence is affected by both kinds of measurement error, it is considered to be the best or most inclusive method of estimating (complete) reliability.

From the preceding it is clear that statements such as "(r)eliability was estimated by means of Cronbach's coefficient alpha" may be misleading to the extent that they suggest that this coefficient has been chosen among several interchangeable methods to estimate reliability. It may also be misleading to state that there are different kinds of reliability (cf. heading on p. 91 of Anastasi & Urbina, 1997) if these different kinds are thought to refer to different methods to estimate the same error in Equations (2) and (4). Rather, it would be more appropriate to state that there are different kinds of measurement error that may reduce the size of the reliability coefficient. These kinds of measurement error are not interchangeable or substitutive, but cumulative. Neither the test-retest method nor the undelayed parallel or the internal-consistency methods are capable of estimating both kinds of measurement error simultaneously and either one of these methods overestimates complete reliability.

Recent proposals to estimate reliability inclusively

Despite the common knowledge about the multifaceted nature of measurement error, the Hogan et al. (2000) study suggests that the majority of reliability studies have resorted to internal-consistency methods, particularly coefficient alpha, which is unaffected by transient error. However, the first few years of the 21st century have seen three proposals that have been formulated within the classical test theory tradition to address this situation. The designs proposed by Becker (2000), Schmidt et al. (2003) and Green (2003) are intended to investigate complete reliability in the absence of parallel test forms. However, all of them require that either the full-length test or two parallel halves be administered on two separate occasions. This makes it possible to obtain a coefficient of equivalence and stability (CES) as a (complete) reliability estimate which is susceptible to both specific-factor error and transient error. These methods also allow for the computation of a coefficient of equivalence (CE) as a (partial) estimate that is affected by specific-factor error only. If one subtracts the complete estimate (which may be expected to be smaller as it is affected by both specific-factor error and transient error) from the partial estimate, an estimate of the proportion of transient error variance is obtained.

Becker (2000) suggested a *staggered equivalent (Rulon-Flanagan) split-half procedure* that requires that the test be split into two strictly parallel test halves. Becker reviewed several approaches to optimise equivalence for such test halves and then expressed his preference for a factor analysis of the items and their assignment to the two halves in terms of the size of their loadings (on the general factor), their means and standard deviations. The test participants are divided into two groups and on the first occasion the two groups complete different halves, and on the second occasion each group takes the half not done on the first occasion. (This counter-balancing is intended to control for possible order effects.) This operation results in four sets of item data. A CE estimate for the full-length test is obtained by (i) computing coefficient alpha for each of the four sets of item data separately, (ii) taking the mean over all four estimates so obtained (which then reflects the coefficient of equivalence for a half-test), and (iii) stepping this mean up by a factor of 2 by means of the Spearman-Brown formula. A CES estimate for the full-length test is determined by (i) calculating, for each group separately, either coefficient alpha or the Rulon-Flanagan split-half reliability for the combination of the test halves completed on different occasions, and then (ii) taking the mean of the reliabilities so obtained for the two groups.

Schmidt et al.'s (2003) proposal could be labelled a *staggered equivalent (adjusted Spearman-Brown) split-half procedure*. Although it also makes provision for the situation where two parallel test forms are indeed available, the more common situation where there is only one test form will be considered here. Their procedure also requires that the test be divided into two strictly parallel test halves and that these two halves are administered on separate occasions. Their CE estimate is compiled in the same manner as in the case of Becker (2000) except that they drop the counter-balancing requirement. Thus the mean of coefficient alpha for the two half-tests (which yields the coefficient of equivalence, ce, for a half-test) is stepped up by a factor of 2 by means of the Spearman-Brown formula to give the coefficient of equivalence, CE, for the full-length test. However, Schmidt et al. developed the following adjustment to the Spearman-Brown formula to obtain a CES for the full-length test:

$$\text{CES} = [2(ces)]/[(1 + ce)],$$

where *ce* is the coefficient of equivalence, defined earlier, for a half-test, and *ces*, is the coefficient of equivalence and stability of a half-test obtained by correlating the scores obtained on two strictly parallel test halves administered on two different occasions.

Green (2003) proposed a *test-retest alpha* that does not call for the test to be split into two parallel test halves but requires that the same full-length test be administered on two separate occasions. He developed an adjusted formula for Cronbach's coefficient alpha,

$$\text{Test-retest alpha} = [J/(J - 1)][(\Sigma_1\Sigma_2\sigma_{j1},\mathrm{k}_2)/(\sigma_1\sigma_2)],\ j \neq k,$$

where $\Sigma_1\Sigma_2\sigma_{j1},k_2$ is the sum of the covariances of every item performed on the first occasion and every other item performed on the second occasion. As in the case of the regular coefficient alpha, the development of Green's test-retest alpha assumes that the items are essentially tau-equivalent. Just as in the case of the regular coefficient alpha, the numerator of the ratio on the rightmost side of the preceding equation is equal to the sum of all the off-diagonal entries of an item variance-covariance matrix but in the present case, the columns of this matrix refer to the items administered on the first occasion and the rows represent the (same) items administered on the second occasion. As in any variance-covariance matrix, there are $J(J - 1)$ covariance terms in this matrix where half of them are in the upper right triangle and the other half are in the lower left triangle. However, in the present case the one half is no longer a mirror image of the other, as in the one half you would find, for example, the covariance of item *j* administered on the first occasion, and item *k* administered on the second occasion, whereas the corresponding entry in the other half would be the covariance of item *j* administered on the second occasion, and item *k* administered on the first occasion and these two covariances need not be equal. The $\sigma_1\sigma_2$ in the preceding equation is simply the product of the standard deviations of the scores obtained on the two administrations of the test.

The CES estimates in the above procedures rely on the covariances of test halves or items pairs the members of which have been administered on different occasions. As a result, they are less affected by memory effects than is the test-retest reliability coefficient. None of them requires the time-consuming construction of strictly parallel test forms as does the conventional coefficient of equivalence and stability. The procedures proposed by Becker (2000) and Schmidt et al. (2003), however, do require the complicated procedure of dividing a test into two strictly parallel halves.

EMPIRICAL EVIDENCE ON THE RELATIVE SIZE OF TRANSIENT ERROR

Of course, if the proportion of transient-error variance is negligible, it makes little sense to harp on the need to estimate such error. However, available evidence suggests that transient error variance cannot be dismissed lightly. The procedures reviewed in the preceding section have made comparisons possible between a CES estimate (susceptible to all three kinds of measurement error) and a CE estimate (subject to specific-factor and random response error only). As a result, the proportion of transient error variance can be obtained by subtracting the former from the latter. This proportion divided by the CES estimate, multiplied by 100, gives the percentage by which the partial reliability estimate overestimates the complete reliability estimate. Becker (2000) administered the Buss-Perry Aggression Questionnaire on two test occasions with a five-day interval in between them. The CE and CES estimates were 0,791 and 0,777, respectively, for the Anger scale for women. This corresponds to a proportion of transient error variance of only 0,791 – 0,777 = 0,014, which means that the partial estimate overestimates the complete estimate by a percentage of only 100(0,014/0,777) = 1,80. However, in the case of the Hostility scale of the same questionnaire for men the corresponding coefficients were 0,809 and 0,679, respectively, which amounts to a proportion of transient error variance of 0,130 and an overestimate of 19,15 %.

Schmidt et al. (2003) studied cognitive, personality and affective measures administered approximately one week apart. They found that transient error was present in all three of these domains and that it was particularly potent in the domain of affective traits. For example, the positive affectivity measure of the Positive and Negative Affect Schedule of Watson, Clark and Tellegen yielded a CE estimate of 0,82, whereas its CES estimate was only 0,63. The resulting estimate of the proportion of transient error variance was 0,19, which translated into an overestimate of 30,16 %. The corresponding reliability coefficients for the negative affectivity measure of Diener and Emmons' Affect-Adjective Scale were 0,90 and 0,69, respectively, yielding a transient error estimate of 0,21 and an overestimate of 30,43 %. Contrary to their expectations, these authors found that the proportion of transient error variance for a measure in the cognitive domain, namely, the Wonderlic Personnel Test, was not less than that for measures of broad personality traits. For example, the proportion of transient error variance in both the Wonderlic and Rosenberg's Self-Esteem Scale was found to be 0,05 with percentages of overestimation of 6,76 and 6,33, respectively.

Schmidt et al. (2003) used Monte Carlo simulation methodology to generate sampling distributions of among others the estimated proportions of transient error variance. As the standard deviations of these sampling distributions represented the standard errors of these indices, confidence intervals could be established. In none of the examples in the preceding paragraph, except for the Rosenberg scale, did the 90% confidence interval capture a value of zero, which in statistical null hypothesis testing would have meant a failure to reject the null hypothesis of a proportion of zero transient error at the 5 % significance level (one-tailed).

The mean coefficient alpha for Green's (2003) two administrations of a four-item Emotional Expression Scale with a four-week interval was equal to (0,861 + 0,931)/2 = 0,896 and the test-retest alpha was 0,735. In terms of the methodology used above, the proportion of transient error variance and the percentage overestimation were therefore 0,161, and 21,90, respectively.

COMPARISON WITH GENERALISABILITY THEORY

The reliability estimation methods of classical test theory (test-retest, parallel-forms with or without an extended interval, split-half, and coefficient alpha) are aimed at arriving at a reliability coefficient with no interest in differentiating between the various sources of measurement error (transient, specific-factor, etc.) that may attenuate the obtained reliability coefficient. Generalisability theory, by contrast, focusses on the separate estimation of the various sources of error variation (in a measurement procedure), called facets, such as test occasion, test items and their respective interactions with participants. The sampled elements of any of these facets are called conditions. Thus test items may constitute the conditions of the item facet, where the composites of items now conform to randomly parallel tests rather than the strictly parallel tests of classical test theory. Typically the test user does not wish the test scores obtained to be relevant only to the occasion on which the test was administered, or only to the particular set of items contained in the test form that was used. Instead, he or she would like to obtain comparable results if the measurement procedure was performed on any other equally acceptable occasion, or if another set of equally acceptable items was applied. This means that the test user wishes to generalise his or her test results to a large universe of conditions of which only a sample from each facet was used in his or her particular application.

The main objective of a *generalisability study* is to simultaneously estimate the variance (components) associated with test participants, with the conditions of each of the respective facets and with the interactions between these sources. Ideally this calls for a design in which each condition of every relevant facet is completely crossed with each condition of every other facet and with each participant. It uses random-effects (and mixed-effects) analysis of variance to separately estimate the variance (component) associated with each of the resulting main and interaction effects. The variance component for the main effect due to participants corresponds to the true-score variance of classical test theory, whereas all the other variance components reflect some or other kind of error variance. For example, if every participant in a group of participants completes every item on each of two or more occasions, generalisability theory allows for the estimation of each of the variance components in the following extension of Equation (2):

$$\sigma^2_{\text{observed}} = \sigma^2_{\text{participants(p)}} + \sigma^2_{\text{occasion(o)}} + \sigma^2_{\text{items(i)}} + \sigma^2_{\text{po}} + \sigma^2_{\text{pi}} + \sigma^2_{\text{oi}} + \sigma^2_{\text{residual}}, \qquad (6)$$

where $\sigma^2_{\text{participants(p)}}$ corresponds to the true-score variance of classical test theory; $\sigma^2_{\text{occasion(o)}}$ is the variance attributable to test occasions; $\sigma^2_{\text{items(i)}}$ is the variance due to items; σ^2_{po} is the variance associated with the participant-occasion interaction (i.e., transient error), σ^2_{pi} is the variance ascribed to the participant-item interaction (i.e., specific-factor error); σ^2_{oi} is the item-occasion interaction variance and $\sigma^2_{\text{residual}}$ is the residual error variance.

The data collection design implied by Green's (2003) test-retest procedure conforms exactly to such a completely-crossed three-factor design with one observation per cell, so that it allows for each of the variance components in Equation (6) to be estimated. Tabel 1 is the result of the application of the procedures described by Shavelson and Webb (1991) on the data Green used for computing his test-retest alpha and which are given in an appendix to his article. Unlike the reliability coefficient which, as a correlation coefficient, is expressed in a universal metric, the results of a generalisability study may be reported in terms of the percentages of the total variance that are accounted for by the respective sources of variation. The largest variance component in Table 1, namely, that for test participants (56,25%), represents the variance component for universe (true) scores. Next highest is the variance component for the residual error (25,87%) which represents the variance due to the three-way interaction plus random response error. The variances of zero for items and occasions are to be expected as under the assumption of parallelism in Green's example, these sources should show no variation. The variance component for the participant-occasion interaction (transient error) is higher than that for the participant-item interaction (specific-factor error).

TABLE 1
ESTIMATED VARIANCE COMPONENTS

Source of Variation	df	Mean Square	Est. Variance Component	%
Persons (p)	39	4,4218	0,4630	56,25
Items (i)	3	0,0166	0	0
Occasions (o)	1	0,0125	0	0
p × i	117	0,2966	0,0418	5,08
p × o	39	0,3643	0,1053	12,80
i × o	3	0,1125	0	0
p × i × o,e	117	0,2129	0,2129	25,87

In the case of Becker (2000) and Schmidt et al. (2003) a different set of items is administered on different occasions so that in their designs items are nested within occasions. As a result, their designs do not allow for the estimation of the variance due to the item-occasion interaction, which is of little interest in any case, or for the estimation of the variance component for the participant-item interaction separately from the residual error variance. In all three procedures reviewed above, the proposed data analysis nevertheless extracts only two estimates of a confounded source of error variation – one in which transient error variance is included and one in which it is excluded. By subtracting the former from the latter in a follow-up analysis, an estimate of transient error variance is obtained. However, none of these procedures provides for the estimation of the specific-factor or residual error variance. If specific-factor error is more detrimental to reliability than is transient error, none of them would be able to detect it. It could be said that these designs underutilise the data that are collected in terms of them.

Generalisability theory has a further advantage in that it provides clear guidelines for the reduction of those sources of error variation that have been shown up by a generalisability study to be unacceptably high. On the basis of the generalisability study results, a prospective user of the measurement procedure may design a *decision study* to estimate the generalisability attendant on his or her intended use of the procedure. This may lead to a decision to sample more heavily from a facet that the generalisability study has revealed to have an unacceptably high error variance. The principle involved here is similar to that which underlies that reformulation of the Spearman-Brown formula that allows one to determine the number of items that is required to reduce specific-factor error variance and, hence, improve reliability satisfactorily. For example, the results of a decision study may reveal that reliability is best improved by administering the same number of items on more occasions (and averaging over occasions) rather than by increasing the number of items. By contrast, the three procedures reviewed above not only are unable to identify the more potent source of error variation but they are also silent about any suggested course of action in situations in which a particular source of error variation is found to be unacceptably high.

For example, in terms of the data in Table 1 one could compute a generalisability coefficient (Shavelson & Webb, 1991), the counterpart of the reliability coefficient of classical test theory, for several possible applications of the measurement procedure in Green's (2003) example. For four items administered on two occasions, the generalisability coefficient is computed to be equal to 0,716. If the number of items is increased by 50 % (so that six items are administered on two occasions), generalisability increases to 0,857. If the number of test occasions is increased by the same percentage (so that four items are administered on three occasions), generalisability is raised to 0,880. With transient error accounting for a greater proportion of error variation than does specific-factor error (cf. Table 1), it is to be expected that generalisability will benefit to a greater extent by an increase in the number of test occasions than by an increase in the number of items.

Finally, unlike classical test theory, generalisability theory also makes it possible to estimate generalisability in contexts where absolute rather than relative decisions are called for. Relative decisions are at stake when the purpose of measurement is to choose a certain number of participants with the highest scores (for the purpose of hiring, promotion or the granting of bursaries, for example). In the case of absolute decisions, tests are used to determine which participants achieve a score above a fixed cut-off value as is typically required in criterion-referenced measurement. An example of this would occur when applicants for a driver's license have to score above a particular minimum value on a test of traffic regulations to qualify for a learner's driver license irrespective of the number of candidates who fail to reach this cut-off score.

Despite the advantages afforded by generalisability theory, the statistically intimidating nature of this extensive approach seems to have thwarted its general acceptance in psychological research. In the Hogan et al. (2000) survey referred to above, not a single generalisability study had been identified. However, their review probably failed to do justice to the popularity of this methodology, as it has seemed to be more useful and popular in educational research. It is particularly well suited to evaluating reliability in performance assessment (such as that involved in outcomes-based research) where the raters of learners' performance represent an important source of error variance (cf. Cronbach, Linn, Brennan & Haertel, 1997).

As indicated earlier, a proper understanding of generalisability theory requires a sound knowledge of multi-factor analysis of variance and not the more common fixed-effects variety but the lesser known random-effects model. Applied psychologists possibly may argue that they use tests in a much simpler context than that catered for by generalisability theory. For example, they may argue that their use of multiple measures of the same construct reduces the magnitude of specific-factor error variance. The appeal of Becker (2000), Schmidt et al. (2003) and Green's (2003) procedures lies in their being developed within the simpler classical test theory tradition and their results being expressed in the same metric as the reliability coefficient. This possibly may improve these procedures' chances of being accepted more widely among psychologists who are concerned about transient error.

DISCUSSION

One of the ironies of psychological test theory and practice is that, conceptually, reliability implies consistency across time – notice the terms *retested* and *repeated administration* in the definitions in the first section – whereas coefficient alpha, the reliability estimation procedure of overwhelming choice, relies on a single test occasion only. This popularity can only be justified if the various reliability estimation methods, including coefficient alpha, may be viewed as interchangeable for the purposes of estimating (the same) measurement error. However, the empirical evidence reviewed in this article suggests that the virtually universal use of coefficient alpha as method of reliability estimation comes at a price. If scores on a test are susceptible to occasional fluctuations in participants' mood or motivation, such transient error can be estimated only if the test or part of it is administered on more than one occasion. In the absence of such retesting all estimation methods will result in overestimates.

One of the tenets of generalisability theory is that what constitutes measurement error and what represents true-score variance depends on the purpose of measurement. Suppose the test includes all possible items that could be formulated so that no parallel test form is feasible and generalising over test forms is, therefore, not called for. In such a case, specific-factor variance should not be treated as error variance but should be incorporated into true-score variance. However, instances where no generalisation over time is intended must be relatively rare. Such situations would tend to run counter to the very notion that led to the conceptualisation of psychometric reliability in the first place.

Although researchers who use tests developed earlier by others may be excused for relying on the partial reliability estimates provided by the original test developer, the latter party cannot similarly be exempted. Test users nevertheless should be aware of the inadequacy of partial estimates and they should clearly identify the estimation method that the original test developer had used.

If specific-factor error and transient error are the only sources of error that are suspected of being present, and if there is no interest in estimating them separately, the above procedures for estimating complete reliability may be sufficient. Although generalisability may present a more comprehensive procedure for estimating error components separately, this approach itself is unable to compensate for the failure to apply a test, or part of it, on different occasions if transient error is present. As far as the estimation of complete reliability is concerned, it is indeed a case of no psychometric gain without test readministration pain.

REFERENCES

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5* (3), 370-379.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Chicago: Holt, Rinehart & Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L. & Haertel, E.H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373-399.

Feldt, L.S & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education.

Gleser, G.C., Cronbach, L.J. & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*, 395-418.

Green, S.B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, *8* (1), 88-101.

Hogan, T.P., Benjamin, A. & Brezinski, K.L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523-531.

Hoyt, C. (1941). Test reliability obtained by analysis of variance. *Psychometrika*, *6*, 153-160.

Jackson, R.W.B. & Ferguson, G.A. (1941). *Studies on the reliability of tests*. Toronto: University of Toronto.

Li, H., Rosenthal, R. & Rubin, D.B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1* (1), 98-107.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Schmidt, F.L. & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

Schmidt, F.L., Le, H. & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8* (2), 206-224.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, Ca.: Sage.

Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, D.C.: American Council on Education.

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 560-620). Washington, D.C.: American Council on Education.

Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.