PSYCHOMETRIKA—VOL. 69, NO. 1, 147–157
MARCH 2004
NOTES AND COMMENTS

RELIABILITY AS A FUNCTION OF THE NUMBER OF ITEM OPTIONS DERIVED FROM THE “KNOWLEDGE OR RANDOM GUESSING” MODEL

ROBERT G. MACCANN

MEASUREMENT AND RESEARCH SERVICES
BOARD OF STUDIES NSW, SYDNEY, AUSTRALIA

For (0, 1) scored multiple-choice tests, a formula giving test reliability as a function of the number of item options is derived, assuming the “knowledge or random guessing model,” the parallelism of the new and old tests (apart from the guessing probability), and the assumptions of classical test theory. It is shown that the formula is a more general case of an equation by Lord, and reduces to Lord’s equation if the items are effectively parallel. Further, the formula is shown to be closely related to another formula derived from Lord’s randomly parallel tests model.

Key words: guessing (tests), test reliability, distractors, test length, error of measurement

1. Introduction

For dichotomously scored multiple choice tests, researchers have long attempted to relate test reliability to the number of options per item. In theory, an increase in the number of options should reduce the probability of guessing correct answers, which should increase test reliability. In early theoretical work, Lord (1944) derived a formula relating test reliability to the number of options from a consideration of the phi coefficient between items. He assumed the “knowledge or random guessing model,” equal item difficulties and equal item intercorrelations. Later, Horst (1954) used Carroll’s (1945) work to relate test reliability to the number of options for a test where the only error considered was that from guessing, giving “immediate retest reliability.” Mattson (1965), using Lord’s (1957, 1959) randomly parallel tests model, constructed a table showing reliability as a function of option numbers, but his table is based on an unobservable quantity, the mean proportion of answers known in the hypothetical population of items. Basing his approach on empirical data, Ebel (1969) took the interval between the maximum possible score and the expected chance score and, assuming the mean was at the midpoint and the standard deviation was one sixth of the width, employed KR21 to relate test reliability to option numbers. Later theoretical work (Grier, 1975; Lord, 1977) considered the option number that maximised test reliability, assuming that for fixed testing time, the product of item and option numbers would be constant, an assumption Lord acknowledged was likely to be false for many or most item types.

This paper extends this work in deriving a formula relating test reliability to the number of options, and showing the relationships between this formula and the work of Lord (1944) and Mattson (1965). To place this development in context, suppose test developers wish to increase test reliability, while retaining the same number of items, and desire an approximate reliability estimate were the number of options to be increased from three to four. One approach would involve writing an additional distractor to each item of an existing test and trialling it on a large sample similar to the examinee population. Alternatively, an approximate reliability estimate could be obtained from theory. In developing this theory, it is not necessary to assume that the individual test items be parallel (as Lord, 1944, assumed). What is required is that, when guessing

Requests for reprints should be addressed to: Dr. Robert MacCann, Head, Measurement & Research Services, Board of Studies NSW, GPO Box 5300, Sydney, AUSTRALIA 2001 E-mail: maccann@boardofstudies.nsw.edu.au

0033-3123/2004-1/2001-0939-A $00.75/0

error is disregarded, the new test as a whole should be parallel in the classical test theory (CTT) sense to the old test. This parallelism assumption is a very stringent one and is to be discussed in a later section.

The development below will employ the widely used "knowledge or random guessing model" (Lord and Novick, 1968, pp. 302–305). This model assumes that the students either know the answer to an item, or that they randomly guess with a probability of success equal to the inverse of the number of options. Neither of these assumptions is plausible for all student responses. While some responses to items may involve definite knowledge by students, other responses may involve varying degrees of partial knowledge. In addition, the random guessing assumption implies that all options are equally plausible to someone not knowing the correct answer. This assumption is unlikely to hold for many items. Indeed, it is commonly held that as more distractors are required, the more difficult it is to construct additional distractors with the same degree of plausibility. Although this model cannot take account of a student's partial knowledge or misinformation about the options in an item, it has been widely used because of its simplicity and is the basis for the correction for guessing formula employed by many testing organizations.

This derivation uses the knowledge or random guessing model and results for CTT and parallel tests to establish an initial set of results, from which expressions for the mean, variance, and reliability of the original test are found. These expressions are then used in determining a formula for the reliability when *the number of options is changed.* It is then shown that this formula reduces to Lord's formula for items that are effectively parallel. It is also shown that in the randomly parallel tests approach, Mattson omitted a nontrivial term in estimating the error variance. By including this omitted term, it is shown that a result can be obtained that is closely related to the formula derived here.

2. The Model

Let X_{if} be a random variable denoting the observed score of student i on form f of an infinite number of parallel forms, each comprising k items. Let X'_{if} be a random variable denoting the actual number of answers *known*, the propensity distribution reflecting errors of measurement in the interaction of person i with form f. On the remaining $k - X'_{if}$ items to which the student does not know the correct answer, student i is assumed to randomly guess on each with a probability of a. These $k - X'_{if}$ attempts with a constant guessing probability constitute a set of Bernoulli trials. Thus the probability distribution for the number of successful guesses of student i on form f will be the binomial distribution with an expected score of $(k - X'_{if})a$ and a variance of $(k - X'_{if})a(1 - a)$. The number of answers known is a random variable, $X'_{i\bullet}$, over forms, with the associated guessing probability distributions for each individual varying over forms.

A model for the observed score of student i on form f can be written as

$$X_{if} = X'_{if} + (k - X'_{if})a + G_{if}. \tag{1}$$

That is, the observed score can be expressed as the number of answers known plus the expected number of answers guessed plus a guessing error.

2.1. *Guessing Error and CTT Results*

As the expected number of answers guessed has been incorporated as a term in the model, the random guessing error, G_{if}, has an expected value of 0, taken over the guessing probability distribution, for a fixed student i and a fixed form f:

$$\mathcal{E}G_{if} = 0. \tag{2}$$

The variance of the guessing probability distribution for student i and form f is given by

$$\sigma^2(G_{if}) = a(1-a)(k - x'_{if}). \tag{3}$$

The guessing error variance over persons for a fixed form f may be found by using Theorem 2.6.2 in Lord and Novick (1968):

$$\sigma^2(G_{\bullet f}) = \mathcal{E}_i \sigma^2(G_{if}) + \sigma^2(\mathcal{E} G_{if}).$$

Applying (2) and (3) and dropping subscripts for persons and forms, the guessing error variance is obtained in terms of the mean of the answers known:

$$\sigma^2(G) = a(1-a)(k - \mu_{X'}). \tag{4}$$

Now on a parallel form of the test, a student may "know" a different number of answers, due to the different sampling of items. Error scores and true scores for the *answers known* are defined below in (5).

$$E'_{if} \equiv X'_{if} - \tau'_i, \quad \text{where } \tau'_i \equiv \mathcal{E}_f X'_{if}. \tag{5}$$

These may be compared to the error and true scores in CTT as defined in (6).

$$E_{if} \equiv X_{if} - \tau_i, \quad \text{where } \tau_i \equiv \mathcal{E}_f X_{if}. \tag{6}$$

The difference between these two error scores is that (5) has had the random guessing error removed, but otherwise incorporates all the sources of error associated with (6).

From (2), the expected value of the guessing error distribution will be zero for for every person i such that τ'_{if} is any specified constant. That is, the regression function $\mathcal{E}_i(G_{if} | \tau'_{if})$ has a constant value of zero. This yields

$$\rho(T'_{\bullet f}, G_{\bullet f}) = 0. \tag{7}$$

A similar argument also yields

$$\rho(G_{\bullet f}, E'_{\bullet f}) = 0. \tag{8}$$

As the true scores of the answers known for a person are identical across the different forms, then (7) also implies

$$\rho(T'_{\bullet f_1}, G_{\bullet f_2}) = 0. \tag{9}$$

Two further results require the assumptions of linear experimental independence, $\mathcal{E}(E'_{if_1} | g_{if_2}) = \mathcal{E}(E'_{if_1})$ for all persons and $\mathcal{E}(G_{if_1} | g_{if_2}) = \mathcal{E}(G_{if_1})$ for all persons, assumptions that will hold if the guessing is random. These imply

$$\rho(E'_{\bullet f_1}, G_{\bullet f_2}) = 0, \tag{10}$$

$$\rho(G_{\bullet f_1}, G_{\bullet f_2}) = 0. \tag{11}$$

With the expected value definition of true score in (5), the parallelism of the forms and the assumption of linear experimental independence for different measurements, the CTT model can be obtained by "construction" rather than assumption (see Lord and Novick, 1968; Zimmerman, 1975, 1976) to give:

$$\mathcal{E}_i E'_{if} = 0, \tag{12}$$

$$\rho(E'_{\bullet f}, T'_{\bullet f}) = 0, \tag{13}$$

$$\rho(E'_{\bullet f_1}, T'_{\bullet f_2}) = 0, \tag{14}$$

$$\rho(E'_{\bullet f_1}, E'_{\bullet f_2}) = 0. \tag{15}$$

2.2. *Parallel Form Results for Alternative True Score Definition*

In this section, certain results holding for parallel forms in terms of the the usual definitions of true and error scores (6), are shown to hold for the *alternative definitions* (5). From (1), (2), (5), and (6), the two sets of true scores for person i are shown to be related:

$$\tau_i = \tau'_i + (k - \tau'_i)a = (1 - a)\tau'_i + ka. \tag{16}$$

Note from (16) that the orthodox true score is larger than the true score of the known answers, the former being inflated by the expected number of items correctly guessed.

From (1), (5), (6), and (16) it may be shown that $E_{if} = (1 - a)E'_{if} + G_{if}$, and hence using (8) that $\sigma^2_E = (1 - a)^2\sigma^2_{E'} + \sigma^2_G$. If the parallel form relationship $\mu_{X_1} = \mu_{X_2}$ holds, then from (1), (2), and (4), $\sigma_{G_1} = \sigma_{G_2}$, and thus (17) follows:

$$\begin{array}{rll} \text{If} & \sigma_{E_1} = \sigma_{E_2}, & \text{then } \sigma_{E'_1} = \sigma_{E'_2} = \sigma_{E'}. \end{array} \tag{17}$$

$$\begin{array}{rll} \text{From (16), if} & \sigma_{T_1} = \sigma_{T_2}, & \text{then } \sigma_{T'_1} = \sigma_{T'_2} = \sigma_{T'}. \end{array} \tag{18}$$

$$\begin{array}{rll} \text{From (16), if} & \mu_{T_1} = \mu_{T_2}, & \text{then } \mu_{T'_1} = \mu_{T'_2} = \mu_{T'}. \end{array} \tag{19}$$

2.3. *Mean, Variance, and Reliability of Original Test*

From (1) and (5), dropping the persons and forms subscripts, gives

$$X = (1 - a)T' + ka + G + (1 - a)E'. \tag{20}$$

From (20) and using (2) and (12), the mean of X may be written as

$$\mu_X = (1 - a)\mu_{T'} + ka. \tag{21}$$

Also from (20), and using (4), (5), (7), (8), (12), and (13), the variance of X is

$$\sigma^2_X = (1 - a)^2\sigma^2_{T'} + a(1 - a)(k - \mu_{T'}) + (1 - a)^2\sigma^2_{E'}. \tag{22}$$

Applying (20) to parallel forms, X_1 and X_2, and from (9)–(11), (14), and (15), the covariance may be written $\text{Cov}(X_1, X_2) = (1 - a)^2\text{Cov}(T'_1, T'_2)$, which from (18) and the parallel result, $\sigma_{X_1} = \sigma_{X_2} = \sigma_X$, yields the reliability coefficient:

$$\rho_{XX'} = (1 - a)^2\frac{\sigma^2_{T'}}{\sigma^2_X}. \tag{23}$$

2.4. *Reliability for a Parallel Test with a Different Number of Options*

Another parallel test, Z, derived from the same test specifications blueprint as X, but with a different number of options per item, has a guessing probability, b. A similar equation to (23) holds for Z. Combining this with (23) gives

$$\rho_{ZZ'} = \frac{(1 - b)^2}{(1 - a)^2}\frac{\sigma^2_X}{\sigma^2_Z}\rho_{XX'}. \tag{24}$$

An expression for the ratio of the variances in (24) is sought. From the equivalent of (22) for Z, and employing (17)–(19), gives

$$\sigma^2_{T'} + \sigma^2_{E'} = \frac{1}{(1-b)^2}\sigma^2_Z - (k - \mu_{T'})\frac{b}{1-b}. \tag{25}$$

A corresponding expression for X exists, which when combined with (25) to eliminate the left hand side, gives:

$$\frac{\sigma^2_Z}{\sigma^2_X} = \frac{(1-b)^2}{(1-a)^2}\left[1 + \frac{(b-a)(1-a)}{(1-b)}\frac{(k-\mu_{T'})}{\sigma^2_X}\right]. \tag{26}$$

Substituting (21) into (26) gives

$$\frac{\sigma^2_Z}{\sigma^2_X} = \frac{(1-b)^2}{(1-a)^2}\left[1 + \frac{(b-a)}{(1-b)}\frac{(k-\mu_X)}{\sigma^2_X}\right]. \tag{27}$$

Substituting (27) into (24) gives

$$\rho_{ZZ'} = \frac{\rho_{XX'}}{1 + \dfrac{(b-a)}{(1-b)}\dfrac{(k-\mu_X)}{\sigma^2_X}}. \tag{28}$$

In estimating the guessing probabilities, the commonly used but implausible assumption is made that the probability of guessing is the inverse of the number of options, giving $a = 1/n_1$ and $b = 1/n_2$. Substituting these in (28) gives

$$\rho_{ZZ'} = \frac{\rho_{XX'}}{1 + \dfrac{(n_1-n_2)}{n_1(n_2-1)}\dfrac{(k-\mu_X)}{\sigma^2_X}}. \tag{29}$$

Equation (29) gives the desired result, showing the new reliability as a function of the old reliability and the numbers of options per item. It does not make an assumption about the homogeneity of X—the method chosen for the estimation of $\rho_{XX'}$ would depend on the structure of X (see the procedures in Feldt and Brennan, 1989). From (29) the predicted change in reliability will be greater when the initial number of options is relatively small than when it is large. For example, if $\lambda = (n_1 - n_2)/(n_1(n_2-1))$ and the option number changes from 3 to 4, $\lambda = -0.11$, but if it changes from 10 to 11, $\lambda = -0.01$. Thus, it predicts diminishing returns as options are being added. Secondly, as the test mean approaches the maximum possible score, the effect of the change in options on reliability becomes minimal. This is in accord with expectations, as a difficult test would involve a relatively high degree of guessing and would allow scope for the change in options to affect the reliability. A very easy test, on the other hand, would involve little guessing, rendering the change in options less influential.

3. Reduction to Lord's Equation

If the items are homogeneous in content, and are nearly of equal difficulty, then the KR21 coefficient shown below may be substituted:

$$\frac{k-\mu_X}{\sigma^2_X} = \frac{k}{\mu_X}\left[1 - \frac{k-1}{k}\rho_{XX'}\right]. \tag{30}$$

For this formula to accurately estimate reliability, it requires item tau-equivalance as well as equal item difficulties. But for (0, 1) scored items, equal item difficulties implies equal item variances which, when considered with the equal item true score variances from tau-equivalence, implies equal item error variances, if errors are uncorrelated with true scores. Thus this substitution is justified if the items are effectively parallel. When (30) is substituted into (29), it can be shown that (29) reduces to Lord's (1944) formula. After conversion to the same notation as this paper, Lord's result may be written

$$\rho_{ZZ'} = \frac{\rho_{XX'}}{1 + \dfrac{(n_1 - n_2)}{n_1(n_2 - 1)}\left[1 - \dfrac{(k-1)}{k}\rho_{XX'}\right]\dfrac{1}{\mu_p}}, \tag{31}$$

where μ_p is the average item proportion correct.

To explore the relationship between (29) and (31), consider the case where the reliability of X is estimated by KR20 for both formulas. If the assumption of equal difficulties for each item is satisfied, then one would expect the two formulas to give the same result. If it is not, the formulas should differ. Rearranging the usual formula for KR20 gives

$$\frac{k - \mu_X}{\sigma_X^2} = \frac{(k - \mu_X)}{\sum p(1-p)}\left[1 - \frac{(k-1)}{k}\rho_{XX'}\right]. \tag{32}$$

Using the relations $\sum p(1-p) = k[\mu_p(1-\mu_p) - \sigma_p^2]$ and $\mu_X = k\mu_p$ in (32) and substituting the result in (29) gives

$$\rho_{ZZ'} = \frac{\rho_{XX'}}{1 + \dfrac{(n_1 - n_2)}{n_1(n_2 - 1)}\left[1 - \dfrac{(k-1)}{k}\rho_{XX'}\right]\dfrac{1}{\left[\mu_p - \dfrac{\sigma_p^2}{(1-\mu_p)}\right]}}. \tag{33}$$

A comparison of (31) and (33) shows that they will be equivalent if the variance of the item difficulties is zero. If the latter is not the case, and the reliabilities are estimated by KR20, then (33) will predict a larger reliability increase if the number of options is increased, and a larger reliability decrease if the number of options is decreased.

4. Relationship to Randomly Parallel Tests Approach

Mattson's work is based on Lord's (1957, 1959) model of randomly parallel tests, in which a person's variance error of measurement is given by the binomial formula $SE_i^2 = k\phi_i(1 - \phi_i)$, where ϕ_i is the proportion correct that person i obtains in the hypothetical item population. In seeking an expression for the variance error of measurement over the population of examinees, Mattson simply replaced each proportion correct with the population average to obtain $SE^2 = k\mu_\phi(1 - \mu_\phi)$. However, as Lord (1957) indicates, this estimate of the variance error of measurement should be obtained by averaging over all examinees to give $SE^2 = \mathcal{E}_i k\phi_i(1-\phi_i) = k\mu_\phi(1-\mu_\phi) - k\sigma_\phi^2$. This gives an extra term, involving the variance of the proportion correct in the population of items over the population of examinees. (See also Lord and Novick, 1968, Equation 11.9.4, for the equivalent expression for observed scores.) With this substitution of the correct expression for the average variance error of measurement, it will be shown below that the randomly parallel tests approach *almost* reduces to (29), with the difference being that the resulting equation is a function of the unobservable true score mean, rather than the observed score mean.

For Z, the correct expression for the variance error of measurement in terms of *true scores* is found by multiplying ϕ by the number of items in the previous formula to give $\sigma_{E_Z}^2 = (1/k)[\mu_{T_Z}(k - \mu_{T_Z}) - \sigma_{T_Z}^2]$. Thus the observed variance of Z is given by:

$$\sigma_Z^2 = \sigma_{T_Z}^2 + \frac{1}{k}[\mu_{T_Z}(k - \mu_{T_Z}) - \sigma_{T_Z}^2]. \tag{34}$$

Mattson's Equation 6 implies $\sigma_{T_Z} = (1 - b)\sigma_{T'}$, which, when substituted in (34), gives

$$\sigma_Z^2 = \frac{(k-1)}{k}(1-b)^2\sigma_{T'}^2 + \frac{\mu_{T_Z}(k-\mu_{T_Z})}{k}. \tag{35}$$

By writing a parallel equation to (35) for form X and rearranging gives

$$\sigma_{T'}^2 = \frac{k}{(k-1)}\frac{1}{(1-a)^2}\left[\sigma_X^2 - \frac{\mu_{T_X}(k-\mu_{T_X})}{k}\right]. \tag{36}$$

Substituting (36) in (35) gives

$$\sigma_Z^2 = \frac{(1-b)^2}{(1-a)^2}\left[\sigma_X^2 - \frac{\mu_{T_X}(k-\mu_{T_X})}{k}\right] + \frac{\mu_{T_Z}(k-\mu_{T_Z})}{k}. \tag{37}$$

For form X, Mattson's Equation 6 also implies $\mu_{T_X} = (1-a)\mu_{T'} + ka$. A similar result applies to form Z. Combining these results gives

$$\mu_{T_Z} = \frac{(1-b)}{(1-a)}(\mu_{T_X} - ka) + kb. \tag{38}$$

Substituting (38) in (37) and simplifying gives

$$\frac{\sigma_Z^2}{\sigma_X^2} = \frac{(1-b)^2}{(1-a)^2}\left[1 + \frac{(b-a)}{(1-b)}\frac{(k-\mu_{T_X})}{\sigma_X^2}\right]. \tag{39}$$

Defining reliability as the ratio of true to observed score variance and substituting Mattson's result, $\sigma_{T_Z} = (1-b)\sigma_{T'}$, gives

$$\rho_{ZZ'} = \frac{(1-b)^2\sigma_{T'}^2}{\sigma_Z^2}. \tag{40}$$

Using a parallel expression for form X and combining it with (40) gives

$$\rho_{ZZ'} = \frac{(1-b)^2}{(1-a)^2}\frac{\sigma_X^2}{\sigma_Z^2}\rho_{XX'}. \tag{41}$$

Substituting (39) into (41) and assuming each guessing probability is the inverse of the number of options gives the result for the randomly parallel model:

$$\rho_{ZZ'} = \frac{\rho_{XX'}}{1 + \frac{(n_1-n_2)}{n_1(n_2-1)}\frac{(k-\mu_{T_X})}{\sigma_X^2}}. \tag{42}$$

It can be seen that (42) is identical to (29) except that the mean of the *true scores* of X in the former replaces the mean of X in the latter. In the CTT model with parallel forms, these two terms would be equal. In the model of randomly parallel tests, however, some forms produced by the process may be significantly easier or more difficult than the average form—for example, for an easier form, μ_X would over-estimate μ_{T_X}.

5. Extension to a Generalized Spearman–Brown Formula

If the length of Z is altered by a factor L, then the Spearman–Brown formula gives $\rho_{YY'} = L\rho_{ZZ'}/(1 + (L-1)\rho_{ZZ'})$. Substituting (29) into this formula gives

$$\rho_{YY'} = \frac{L\rho_{XX'}}{1 + (L-1)\rho_{XX'} + \frac{(n_1-n_2)}{n_1(n_2-1)}\frac{(k-\mu_X)}{\sigma_X^2}}. \tag{43}$$

This gives the generalized Spearman–Brown formula, which predicts the new reliability after a change in the length of the test and a change in the number of options. Apart from the assumptions required in changing the number of options, this formula assumes that the new item set resulting from the change in length is parallel to the existing set (except for length).

6. Application to Data

Equations (29) and (31) were applied to five tests containing a 4-option multiple choice component from statewide achievement examinations in NSW, Australia. These were English (20 items), Mathematics (45 items), and Biology, Chemistry, and Physics (each 15 items). Each test sample had approximately 10,000 respondents. Table 1 shows the means and standard deviations of the original tests plus some predicted reliabilities.

For example, English containing 4 options had a KR20 reliability of 0.704. Using (29), a 3-option version of the test of the same length is predicted to have a reliability of 0.655, and a 5-option version, a reliability of 0.732. For (31), the respective predicted reliabilities are 0.659 and 0.730. For each data set, the predicted difference between (29) and (31) was small.

Now consider the context described earlier where the test developers have a 3-option test and are wondering what the reliability would be if an extra option was written for each item. Such an experiment has been performed by Trevisan, Sax, and Michael (1994), although the focus of their investigation was different to that of this paper. Equations (29) and (31) were applied to the published data in Trevisan et al. for this case with the results shown in Table 2 below.

The KR20 reliability of the 45 item 3-option test was originally 0.65. When an extra distractor was written, the 4-option version was administered to a randomly equivalent group of examinees, giving a resulting reliability of 0.75. Using the Trevisan et al. published mean and standard deviation, the estimated 4-option reliabilities were 0.72 for (29) and 0.70 for (31). In this case, the predicted values have under-estimated the reliability. However, in cases where the additional distractors written are of lower quality than the existing distractors, the equations may be expected to over-estimate the reliability.

TABLE 1.
Reliabilities predicted by equations (29) and (31)

				Observed	Predicted			
					Equation (29)		Equation (31) (Lord's Equation)	
Course	k	Mean	SD	4 option	3 option	5 option	3 option	5 option
English	20	11.976	3.640	.704	0.655	0.732	0.659	0.730
Mathematics	45	24.598	7.086	.834	0.794	0.856	0.800	0.852
Biology	15	11.706	2.330	.628	0.583	0.652	0.588	0.649
Chemistry	15	9.385	3.123	.741	0.692	0.769	0.698	0.765
Physics	15	8.686	2.866	.646	0.589	0.678	0.594	0.675

TABLE 2.
Reliabilities for 4-option test estimated from 3-option test

	Observed		Estimated 4 option	
Summary Statistics 3 option	3 option	4 option	Equation (29)	Equation (31) (Lord's Equation)

Mean = 24.11
SD = 4.88
	.65	.75	.72	.70

Data taken from Trevisan et al., 1994

7. Discussion

For 3 and 4-option parallel tests, one may compare 3-option test reliability to that obtained if the original options are kept and an additional option per item is written. Trevisan et al. (1994) argue that this procedure of creating test forms *incrementally* more closely approximates actual test construction procedures and is a natural comparison to make. An alternative comparison is between a 3-option test and an entirely newly written 4-option test, as would result, for example, if one were comparing last year's test with this year's test. In either case, for the formula to be effective, the new form must be written such that, disregarding guessing error, it is parallel to the old form (i.e., that Equations (17)–(19) hold). Although it is an empirical question as to which scenario would be more effective in producing parallel tests, the incremental approach gives the *same subject matter* across forms per item and this could result in a closer degree of parallelism. The difficulties in creating parallel forms suggest that in practice, the prediction is likely to be very approximate and that there will be many practical situations where the prediction will be inadequate. On the other hand, the experimental alternative of trialling an existing test with an additional distractor written for each item has its own practical problems (particularly if the original test items cannot be kept secure after their administration) that may result in poor estimation.

Although it may be difficult in practice to satisfy the parallelism requirement, one may question how such a formula can be derived without it. A distinction may be drawn between tests that are already in existence and hypothetical tests that are yet to be built. In the former, the statistics describing the tests are *known* and thus are available to be used in formulas. However, for a test that is yet to be built, its exact properties are unknown. When it is built and administered to examinees, it may turn out to be only essentially tau-equivalent to the existing test, with a slightly different mean and error variance. These differences are not a deliberate feature of the test construction process and cannot be predicted in advance. It could be argued that in requesting a reliability estimate for a 4-option version of a 3-option test, the developers are implicitly requiring the condition of "all other things being equal." That is, if a new test could be built that was like the previous test in all respects, with the only difference being the number of options, what would be the change in reliability? If the request implied a model other than parallelism, such as tau-equivalence, then the estimated difference in reliabilities would become a function of something additional to the change in the number of options (namely, the difference in the error variances), which is not what is logically desired, as this difference is an unknown and unpredictable quantity.

In other approaches to this problem, Lord (1944) assumed that the new test was parallel to the old one (apart from the number of options). Lord's derivation was primarily based at the item level, where his equations assume that each new item was parallel with the old item, apart from the guessing probability associated with the different number of options. Further, he assumed that all items in the existing test were of equal difficulty and that all item intercorrelations were equal, thus assuming parallel items in the existing test. He also assumed the "knowledge or random guessing model." Therefore, his derivation was considerably more restrictive than the one derived here.

The randomly parallel tests approach is not sufficient to provide a formula in terms of observable values, if only a single existing test is available, as the equation requires the mean of the true scores. Both the main development in this paper, and the randomly parallel tests approach, enable Equation (42) to be derived. However, the former, based on the assumption of parallel tests, enables one to substitute the observed mean as a suitable estimate for the true score mean. In the latter, the existing test, being formed by the random selection of items, may be harder or easier than average, giving an observed score mean that may underestimate or overestimate the true score mean. If data from a series of randomly parallel tests were available, then an estimate derived from all these means could be used. The randomly parallel tests approach also uses the “knowledge or random guessing model” and thus also suffers from the limitations of this model.

In estimating the reliability of X in Equation (29), if the original test can be divided into a number of parts that are essentially tau-equivalent, then Cronbach’s alpha could be used. Novick and Lewis (1967), assuming *uncorrelated* errors between the parts, showed theoretically that without essential tau-equivalence, alpha tends to underestimate the reliability, while Zimmerman, Zumbo, and Lalonde (1993) showed this through computer simulation. However, if the parts have correlated errors, then Zimmerman et al. demonstrated that alpha may give an inflated estimate. Komaroff (1997) investigated the simultaneous violation of essential tau-equivalence and uncorrelated errors, concluding that alpha is sensitive to these opposing biases, and may either underestimate or overestimate reliability, depending on the strength of each bias. For simulated data, Raykov (2001) showed the inadequacy of alpha for high degrees of tau-equivalence violation and advocated a covariance structure model which gave accurate reliability estimates. If the parts correspond to individual items which are scored (0, 1), then Cronbach’s alpha reduces to KR20. Feldt and Brennan (1989) point out that this dichotomous scoring inevitably leads to violation of tau-equivalence, but state that KR20 has not been found to seriously underestimate split halves coefficients for tests of reasonably homogeneous content.

Equation (29) is perhaps most effective when used to span only one option level (e.g., from three to four options). However, in practice, a point is likely to be reached where even a span of one option level gives poor prediction: In predicting from five to six options, apart from the diminishing returns predicted by the formula, the latter options may be so ineffective in practice that the actual reliabilities for five- and six-option tests may be virtually the same. One may speculate that the formula is possibly most effective in a narrow range, perhaps in the range from three to five options using only a one-option span. Even so, its usefulness would rely heavily upon the writing of the high quality plausible distractors on which the theory depends.

References

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, *X*, 1–19.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, *29*, 565–570.

Feldt, L.S. & Brennan, R.L. (1989). Reliability. In R.L. Linn (ed.) Educational Measurement (3rd ed., pp. 105–147). New York: American Council on Education; Macmillan.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement*, *12*, 109–113.

Horst, P. (1954). The estimation of immediate retest reliability. *Educational and Psychological Measurement*, *14*, 705–708.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on Coefficient alpha. *Applied Psychological Measurement*, *21*, 337–348.

Lord, F. M. (1944). Reliability of multiple choice tests as a function of choices per item. *Journal of Educational Psychology*, *35*, 175–180.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510–521.

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, *19*, 233–239.

Lord, F. M. (1977). Optimal number of choices per item—a comparison of four approaches. *Journal of Educational Measurement*, *14*, 33–38.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. *Educational and Psychological Measurement*, *25*, 727–730.

Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*, 315–323.

Trevisan, M.S., Sax, G., & Michael, W.B. (1994). Estimating the optimum number of options using an incremental option paradigm. *Educational and Psychological Measurement*, *54*, 86–91.

Zimmerman, D.W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*, 395–412.

Zimmerman, D.W. (1976). Test theory with minimal assumptions. *Educational and Psychological Measurement*, *36*, 85–96.

Zimmerman, D.W., Zumbo, B.D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*, 33–49.

Manuscript received 5 OCT 2001
Final version received 18 MAR 2003