Two Prophecy Formulas for Assessing the Reliability of Item Response Theory-Based Ability Estimates

Nambury S. Raju and T. C. Oshima

Educational and Psychological Measurement 2005; 65; 361
DOI: 10.1177/0013164404267289

TWO PROPHECY FORMULAS FOR ASSESSING THE RELIABILITY OF ITEM RESPONSE THEORY–BASED ABILITY ESTIMATES

NAMBURY S. RAJU
Illinois Institute of Technology

T. C. OSHIMA
Georgia State University

Two new prophecy formulas for estimating item response theory (IRT)–based reliability of a shortened or lengthened test are proposed. Some of the relationships between the two formulas, one of which is identical to the well-known Spearman-Brown prophecy formula, are examined and illustrated. The major assumptions underlying these formulas are outlined and discussed. Both prophecy formulas appear to provide comparable estimates of reliability in the IRT context as well as in the classical test theory context.

Keywords: *reliability; Spearman-Brown prophecy formula; item response theory; classical test theory*

One of the major advantages of item response theory (IRT) over classical test theory (CTT) is the ease with which the standard error of measurement *(SEM)* associated with a given ability level can be assessed (Hambleton & Swaminathan, 1985; Lord, 1980; Wright & Stone, 1979). In IRT, the square of the *SEM* is inversely equal (asymptotically) to the sum of item information *(I)* functions (Hambleton & Swaminathan, 1985). Both the *SEM* and *I* vary as

Portions of this article were presented at the annual meeting of the American Educational Research Association in 1991 in Chicago. The authors would like to express their appreciation to Claudia P. Flowers for her assistance with the data analysis and to Jeffrey A. Slinde, Frank L. Schmidt, and two anonymous reviewers for their comments on an earlier version of this article. Correspondence should be sent to Nambury S. Raju, Institute of Psychology, Illinois Institute of Technology, Chicago, IL 60616-3793; e-mail: raju@iit.edu.

Educational and Psychological Measurement, Vol. 65 No. 3, June 2005 361-375
DOI: 10.1177/0013164404267289

a function of examinee ability (θ), thus resulting in a separate *SEM* for each ability level.

The concept of the reliability of a test for a group or population of examinees, which plays a significant role in CTT, has received minimal attention thus far in the IRT area. As experience with IRT grows, an overall measure of the degree of accuracy of ability estimates, in addition to the localized *SEM*s, may be needed to summarize the data for a group of examinees. An average *SEM* may not be a very useful construct for this purpose because the scale on which the *SEM* is expressed will not readily lend itself to an easy interpretation of the average *SEM*. On the other hand, a measure of reliability could serve a useful purpose in providing such an overall measure of the degree of accuracy of IRT-based ability estimates for a group of examinees. In CTT, the concept of reliability is well known to practitioners, and the fact that it varies between 0 and 1 adds greatly to its ease of interpretation. An IRT-based reliability could also play an equally significant role in the interpretation of test scores as IRT-based test development becomes more widespread. In fact, Lord (1983), Samejima (1994), and Sympson (1980) presented formulas for estimating the reliability of IRT-based ability estimates, which will be described below. Information about some prior concerns pertaining to the concept of reliability within the IRT context may be found in Samejima (1977); Green, Bock, Humphreys, Linn, and Reckase (1984); and Divgi (1989). Also, a recent application of an IRT-based reliability estimation in computer adaptive testing may be found in Nicewander and Thomasson (1999).

Once the reliability of an IRT-developed test is known for a group of examinees, a test developer may want to know what effect the addition of new items to, or the deletion of items from, an existing test would have on the reliability estimate of the expanded or reduced test for the same group of examinees. Such information is generally needed even before the new items are generated because the test developer may want to assess the cost of such an undertaking and determine its payoff prior to making a decision on the investment. It seems that there is a need for new psychometric techniques to assess the effect (on reliability) of adding new items to, or deleting items from, an existing test within the IRT context. That is, there is a need for IRT-based prophecy formulas of the type associated with Spearman and Brown in CTT. The purpose of this article is to develop two prophecy formulas for estimating the reliability of a test (for a group of examinees) that is either shortened or lengthened.

Definition of Reliability

Let $\hat{\theta}_{sx}$ denote an estimate of ability (θ_{sx}) for examinee *s* with test *x*. Let the relationship between $\hat{\theta}_{sx}$ and θ_{sx} be defined as

$$\hat{\theta}_{sx} = \theta_{sx} + e_{sx}, \tag{1}$$

where e_{sx} is the measurement error for examinee s on test x. It should be noted that throughout this article, the theta (θ) and its estimate ($\hat{\theta}$) will have at least one subscript, x or y, to indicate that they are related to items in test x or y; the same notation also holds true for the measurement error. Because the metric underlying the theta scale is arbitrary (Hambleton & Swaminathn, 1985; Lord, 1990), both the theta and its estimate may be defined differently for different tests. Also, different computer programs for IRT calibration may impose different metrics on the theta scale. So a subscript, x or y, is used to emphasize the fact that a particular test or a set of items may be involved in defining the metric for theta and its estimate. Standard errors associated with theta estimates are likely to vary as a result of the number and type of items contained in a test.

Let us now assume that $\hat{\theta}_{sx}$ is an unbiased estimate of θ_{sx}; that is, $E\left(\hat{\theta}_{sx}\right) = \theta_{sx}$, where E is the expectation operator. Then the variance of $\hat{\theta}_{sx}$ over examinees can be written as

$$\sigma^2_{\hat{\theta}_x} = \sigma^2_{\theta_x} + \underset{s}{E}\left(\sigma^2\left(e_{sx}|s\right)\right) = \sigma^2_{\theta_x} + \underset{s}{E}\left(SEM^2_{sx}\right). \tag{2}$$

The expectation term in the middle part of this equation is the variance of measurement error for examinee s on test x and is, therefore, simply the square of the *SEM* for examinee s. In the tradition of CTT, the reliability of theta estimates for a group of examinees may be expressed as

$$\rho_{\hat{\theta}_x\hat{\theta}_x} = \frac{\sigma^2_{\theta_x}}{\sigma^2_{\hat{\theta}_x}} = \frac{\sigma^2_{\hat{\theta}_x} - \underset{s}{E}\left(SEM^2_{sx}\right)}{\sigma^2_{\hat{\theta}_x}}. \tag{3}$$

Equation 3 was previously developed by Samejima (1994) and Sympson (1980) as a measure of an IRT-based reliability of a test for a group of examinees. Lord's (1983) IRT-based reliability formula, which is given as Equation 51 in his article, is very similar to Equation 3, except that Lord's equation contains an additive correction factor. According to Lord, this correction factor is needed because the maximum likelihood estimates of θ are generally biased unless the number of items in the test is extremely large or approaches infinity. Therefore, Equation 3 assumes that an estimate of θ is unbiased.

The variance of $\hat{\theta}_{sx}$ in Equation 3 may be estimated by computing the variance of estimated abilities of examinees in a given group. The expectation of the square of the *SEM* may be estimated by first computing the *SEM* for each examinee (using his or her estimate of theta) and then taking the average of the squared *SEM*s. Because the square of the IRT-based *SEM* is inversely

(and asymptotically) equal to the test information function, the SEM^2 for examinee s can be obtained from

$$SEM_{sx}^2 = \frac{1}{I_{sx}} = \frac{1}{n_x \bar{I}_{sx}}, \tag{4}$$

where I_{sx} and $\bar{I}_{sx}$ are the total and average test information functions, respectively, and n_x is the number of items in test x. Formulas for computing the item and test information functions for the one-, two-, and three-parameter logistic models may be found in Hambleton and Swaminathan (1985) and Lord (1990). Because the test information function is the sum of item information functions, increasing the number of items (that is, adding new items that are similar in content and psychometric quality to the current item pool) will likely increase the total information function, which, in turn, will reduce the *SEM*. Also, according to Equation 3, reducing *SEM* will increase the reliability of a test for a group of examinees. This is also true in CTT.

Equation 3 can yield negative reliabilities if the number of items is small and the average item information function is low. This can happen even when the estimates of theta are unbiased. To illustrate this, let us assume, without loss of generality, that the variance of estimated thetas is one. Then Equation 3 can be rewritten as

$$\rho_{\hat{\theta}_x \hat{\theta}_x} = 1 - \underset{s}{E}\left(SEM_{sx}^2 \right), \tag{5}$$

which, in view of Equation 4, can be further rewritten as

$$\rho_{\hat{\theta}_x \hat{\theta}_x} = 1 - \underset{s}{E}\left(\frac{1}{I_{sx}} \right) = 1 - \frac{1}{n_x} \underset{s}{E}\left(\frac{1}{\bar{I}_{sx}} \right), \tag{6}$$

where the expectation (E) is taken over s. Now, $\rho_{\hat{\theta}_x \hat{\theta}_x} > 0$ if and only if

$$n_x \succ \underset{s}{E}\left(\frac{1}{\bar{I}_{sx}} \right). \tag{7}$$

For example, within the Rasch model (Wright & Stone, 1979), if the average item information is .16, then $n_x$must be at least 7 for the reliability to be greater than 0.

For later use, the reliability formula in Equation 3, in view of Equation 2, can also be expressed as

$$\rho_{\hat{\theta}_x \hat{\theta}_x} = \frac{\sigma_{\theta_x}^2}{\sigma_{\theta_x}^2 + \underset{s}{E}\left(SEM_{sx}^2 \right)}. \tag{8}$$

Although this equation is psychometrically equal to Equation 3, the estimation of IRT-based reliability for a group of examinees is typically done with Equation 3 in practice.

Two Prophecy Formulas for Estimating Reliability

Let n_y be the number of items in a new test (y) derived from test x. It should be noted that the added or deleted items are assumed to be similar in content and psychometric quality to the items that are already in test x.

Prophecy Formula Based on Equation 3

In light of Equation 4, Equation 3 can be rewritten as

$$\rho_{\hat{\theta}_x\hat{\theta}_x} = 1 - \left(\frac{1}{\sigma^2_{\hat{\theta}_x}}\right) E_s\left(\frac{1}{I_{sx}}\right). \tag{9}$$

Rearranging terms, Equation 9 can be rewritten as

$$E_s\left(\frac{1}{\bar{I}_{sx}}\right) = n_x\sigma^2_{\hat{\theta}_x}\left(1-\rho_{\hat{\theta}_x\hat{\theta}_x}\right). \tag{10}$$

Similarly for test y, we have

$$E_s\left(\frac{1}{\bar{I}_{sy}}\right) = n_y\sigma^2_{\hat{\theta}_y}\left(1-\rho_{\hat{\theta}_y\hat{\theta}_y}\right). \tag{11}$$

Because actual data are unavailable for test y, one may want to use the average item information function for test x to estimate the average item information function for test y provided the items in x and y are comparable in terms of content and psychometric quality; that is, the left-hand side of Equation 11 may be approximated by the left-hand side of Equation 10. Therefore,

$$n_x\sigma^2_{\hat{\theta}_x}\left(1-\rho_{\hat{\theta}_x\hat{\theta}_x}\right) = n_y\sigma^2_{\hat{\theta}_y}\left(1-\rho_{\hat{\theta}_y\hat{\theta}_y}\right). \tag{12}$$

Solving for $\rho_{\hat{\theta}_y\hat{\theta}_y}$ in Equation 12, we obtain

$$\rho_{\hat{\theta}_y\hat{\theta}_y} = \frac{\rho_{\hat{\theta}_x\hat{\theta}_x} + kA - 1}{kA}, \tag{13}$$

where $k = \frac{n_y}{n_x}$ and

$$A = \frac{\sigma^2_{\hat{\theta}_y}}{\sigma^2_{\hat{\theta}_x}}. \tag{14}$$

If it is assumed that $A = 1$, then Equation 13 can be rewritten as

$$\rho_{\hat{\theta}_y\hat{\theta}_y} = \frac{\rho_{\hat{\theta}_x\hat{\theta}_x} + (k-1)}{k}, \tag{15}$$

which is a prophecy formula for estimating the reliability of y when only k and the reliability of x are known. As shown in the appendix, the same formula (Formula 15) can also be derived within the CTT context.

Prophecy Formula Based on Equation 8

In view of Equation 4, Equation 8 can be rewritten as

$$\rho_{\hat{\theta}_x\hat{\theta}_x} = \frac{\sigma^2_{\theta_x}}{\sigma^2_{\theta_x} + E_s\left(\frac{1}{n_x \bar{I}_{sx}}\right)}. \tag{16}$$

Rearranging the above equation,

$$E_s\left(\frac{1}{\bar{I}_{sx}}\right) = n_x\sigma^2_{\theta_x}\left(\frac{1-\rho_{\hat{\theta}_x\hat{\theta}_x}}{\rho_{\hat{\theta}_x\hat{\theta}_x}}\right). \tag{17}$$

Similarly for test y,

$$E_s\left(\frac{1}{\bar{I}_{sy}}\right) = n_y\sigma^2_{\theta_y}\left(\frac{1-\rho_{\hat{\theta}_y\hat{\theta}_y}}{\rho_{\hat{\theta}_y\hat{\theta}_y}}\right). \tag{18}$$

As before, let us assume that the average item information functions for x and y are equal; that is, let the left-hand side of Equation 18 equal the left-hand side of Equation 17. Then,

$$n_x\sigma^2_{\theta_x}\left(\frac{1-\rho_{\hat{\theta}_x\hat{\theta}_x}}{\rho_{\hat{\theta}_x\hat{\theta}_x}}\right) = n_y\sigma^2_{\theta_y}\left(\frac{1-\rho_{\hat{\theta}_y\hat{\theta}_y}}{\rho_{\hat{\theta}_y\hat{\theta}_y}}\right). \tag{19}$$

Assuming that $\sigma^2_{\hat{\theta}_x} = \sigma^2_{\hat{\theta}_y}$ and solving for $\rho_{\hat{\theta}_y\hat{\theta}_y}$ in the above equation, one obtains

$$\rho_{\hat{\theta}_y\hat{\theta}_y} = \frac{k\rho_{\hat{\theta}_x\hat{\theta}_x}}{1+(k-1)\rho_{\hat{\theta}_x\hat{\theta}_x}}, \tag{20}$$

which is identical to the Spearman-Brown prophecy formula.

A Comparison of Prophecy Formulas 15 and 20

The goal of the two prophecy formulas is the same: estimating the reliability of an expanded or reduced test when only an estimate of the reliability or the average test information function of the original test is known. Irrespective of which prophecy formula is used, empirical results from both formulas may not be substantially different. Furthermore, when an existing test is expanded, an estimate of reliability from Equation 15 will be (slightly) greater than the estimate of reliability from the Spearman-Brown formula or Formula 20, and the converse is true when a test is reduced. To see these analytical relationships, let us write the ratio of Formula 15 to Formula 20, after simplification, as

$$\frac{Formula(15)}{Formula(20)} = 1+\left(\frac{\left(1-\rho_{\hat{\theta}_x\hat{\theta}_x}\right)^2}{\rho_{\hat{\theta}_x\hat{\theta}_x}}\right)\frac{(k-1)}{k^2}. \tag{21}$$

If k is greater than one (that is, test y is an expanded version of test x), the ratio in Equation 21 will be greater than one, indicating that Formula 15 will yield a higher estimate than Formula 20 or the Spearman-Brown prophecy formula. The converse is true when k is less than one. Despite the derivational differences noted above, the prophecy estimates of reliability from Equations 15 and 20 may be comparable in test-development situations typically encountered in practice. An empirical example is offered as an illustration of the reliability estimates that result from the two prophecy formulas.

An Example

Data from a statewide criterion-referenced test in mathematics were used to illustrate the use of the new prophecy formula as well as the Spearman-Brown formula. This basic skills test consisted of 111 multiple choice items, of which only 80 items were selected for this investigation. The 31 excluded items were the easiest items in the test. Three separate subsets of 20, 40, and 60 items were randomly selected from the pool of 80 items such that the 20-

item test was contained in the 40-item test, which, in turn, was contained in the 60-item test. Finally, the 60-item test was contained in the 80-item test. This selection of items is designed to reflect the common practice when items are added to, or deleted from, an existing test. The sample for the current investigation consisted of 3,000 10th-graders randomly selected from the 63,505 10th-graders who took the criterion-referenced test in 1984.

Prior to the reliability analysis, all four tests were calibrated with the 1-PL, 2-PL, and 3-PL models using the BILOG 3 program (Mislevy & Bock, 1990). The means and standard deviations of raw scores for the 20-, 40-, 60-, and 80-item versions of the test for the 3,000 10th-graders are shown in Table 1. Also shown in this table are the proportional score (total raw score divided by the number of items) means and standard deviations. Table 2 shows the means and variances of estimated thetas from the three (1-PL, 2-PL, and 3-PL) IRT calibrations for the four test-length versions. Also shown in this table are the variances of theta estimates and the average of SEM^2 estimates. Variances of estimated thetas and averages of SEM^2s were directly computed from the BILOG output.

Reliability estimates for each of the four tests were computed with two methods: alpha coefficient (from the CTT framework) and Formula 3 (from the IRT framework). Within the IRT framework, three separate reliability estimates, one for each of the three IRT models, were obtained. Two different scenarios were examined. In Scenario I, the 20-item test was considered as the base or existing test, and the 40-, 60-, and 80-item tests were treated as the expanded tests. In Scenario II, the 80-item test was designated as the existing test, and the remaining three tests were considered the reduced tests. In both scenarios, reliabilities for the expanded and reduced tests were estimated using Formulas 15 and 20. Results for Scenario 1 are shown in Table 3 and for Scenario II in Table 4.

The three IRT-based reliability estimates for the 20-item (base/existing) test in Scenario I (Table 3) are quite similar, hovering around .75. The CTT-based reliability estimate is slightly higher at .81. This trend appears to be generally true also for the remaining three tests. The prophecy estimates (based on Formulas 15 and 20) appear to be generally accurate when compared with the actual (alpha for CTT and Formula 3 for IRT) reliabilities. As expected from Equation 21, the estimates based on Formula 15 are consistently higher than the estimates from Equation 20 (Spearman-Brown formula), but never by more than .02. Both prophecy formulas appear to provide accurate estimates of reliability whether used in the IRT context or in the CTT context. Despite the fact that Formula 20 (Spearman-Brown formula) was originally developed in the CTT context and that Formula 15 was first developed in the IRT context, it is interesting to note that for the current data sets both prophecy formulas appear to provide accurate estimates in either context. Even though the accuracy of the reliability estimates from the two

Table 1
Raw Score Means and Standard Deviations for the 20-, 40-, 60-, and 80-Item Tests and Means and Standard Deviations of Proportional Scores

Test	*M*	*SD*	*M/n*	*SD/n*
20-item test	12.52	4.41	.626	.220
40-item test	25.57	8.34	.639	.209
60-item test	36.50	12.59	.608	.210
80-item test	49.17	16.42	.615	.205

Note. *M/n* = mean of proportional scores; *SD/n* = standard deviation of proportional scores.

Table 2
Means and Variances of Estimated Thetas for the 20-, 40-, 60-, and 80-Item Tests and the Average SEM^2

Test	Mean	Estimated Theta ($\hat{\theta}$) Variance	Average SEM^2
20-Item Test			
1-PL	.015	.801	.198
2-PL	.007	.797	.190
3-PL	.005	.806	.201
40-item test			
1-PL	.039	.951	.123
2-PL	.026	.927	.117
3-PL	.007	.945	.121
60-item test			
1-PL	.036	.970	.081
2-PL	.024	.925	.074
3-PL	.008	.904	.082
80-item test			
1-PL	.039	.994	.064
2-PL	.030	.945	.059
3-PL	.010	.914	.064

prophecy formulas could vary from one data set to another, the estimates themselves will not differ much from each other in view of Equation 21.

The trends in Table 4 are very similar to those noted in Table 3, with one exception. The estimates from Formula 15 are consistently lower than the estimates from Formula 20 (Spearman-Brown formula), but never by more than .04. These lower estimates for Formula 15 are in conformity with Equation 21. Again, for the current data sets, the two prophecy formulas provide accurate estimates of reliability in both (IRT and CTT) contexts. Finally, it should be noted that the same test was calibrated with three different IRT

Table 3
Actual and Estimated Reliabilities When a Test is Expanded

	Items			
	20 (x)	40 (y) (k = 2)	60 (y) (k = 3)	80 (y) (k = 4)
Classical test theory				
Actual (alpha)	.811	.896	.930	.946
Estimate				
Formula 15		.905	.937	.953
Formala 20 (Spearman-Brown)		.896	.928	.945
Item response theory (IRT) (1-PL)				
Actual (Formula 3)	.753	.871	.917	.936
Estimate				
Formula 15		.877	.918	.938
Formula 20 (Spearman-Brown)		.859	.902	.924
IRT (2-PL)				
Actual (Formula 3)	.762	.874	.920	.938
Estimate				
Formula 15		.881	.921	.941
Formula 20 (Spearman-Brown)		.865	.906	.928
IRT (3-PL)				
Actual (Formula 3)	.751	.872	.909	.930
Estimate				
Formula 15		.876	.917	.938
Formula 20 (Spearman-Brown)		.858	.901	.924

Note. x refers to the base/existing test and y refers to the expanded test.

models only for illustrative purposes. We realize that in practice only one IRT model is typically used for calibrating a given test.

Concluding Remarks

The derivation of the two prophecy formulas given in Equations 15 and 20 for estimating IRT-based reliability of a shortened or lengthened test is based on two assumptions: One assumption is common to both formulas, and the second assumption is specific to each formula. The first assumption states that the average item information functions are equal for tests x and y. This assumption may be considered tenable if one observes that y is created by deleting items from x or by adding items to x so that the deleted or added items are similar (in content and psychometric quality) to the items already contained in x.

The second assumption for Formula 15 states that the variance of estimated thetas for test x is equal to the variance of estimated thetas for test y (A = 1). This assumption appears to have justification, given the current practice for

Table 4
Actual and Estimated Reliabilities When a Test is Reduced

	Items			
	20 (*y*) (*k* = .25)	40 (*y*) (*k* = .50)	60 (*y*) (*k* = .75)	80 (*x*)
Classical test theory				
Actual (Alpha)	.811	.896	.930	.946
Estimate				
Formula 15	.782	.891	.927	
Formula 20 (Spearman-Brown)	.813	.897	.929	
Item response theory (IRT) (1-PL)				
Actual (Formula 3)	.753	.871	.917	.936
Estimate				
Formula 15	.745	.873	.915	
Formula 20 (Spearman-Brown)	.786	.880	.917	
IRT (2-PL)				
Actual (Formula 3)	.762	.874	.920	.938
Estimate				
Formula 15	.752	.876	.918	
Formula 20 (Spearman-Brown)	.791	.883	.919	
IRT (3-PL)				
Actual (Formula 3)	.751	.872	.909	.930
Estimate				
Formula 15	.722	.861	.907	
Formula 20 (Spearman-Brown)	.770	.870	.909	

Note. *x* refers to the base/existing test and *y* refers to the reduced test.

defining the theta metrics in IRT calibrations. For example, in the BILOG program (Mislevy & Bock, 1990), thetas are defined to have a mean of zero and a standard deviation of one. That is, in BILOG, estimated thetas have a variance approximately equal to 1.00. That is, two tests, one with 30 items and the other with 50 items, would have the same variance (approximately equal to 1.00) for estimated thetas, unless specified otherwise. The second assumption for Formula 20 states that the variance of thetas is the same for *x* and *y*. Given that the examinees are the same provides sufficient justification for this assumption. It should be noted that a slight variation of these two assumptions is also needed for deriving Formula 15 within the CTT context (see the appendix).

Finally, the two prophecy formulas may generally provide comparable estimates of reliability in both the IRT and CTT contexts. The new prophecy formula (Equation 15) offers an easy-to-use alternative to the Spearman-Brown formula (Equation 20). Despite all this, we feel that there is still a great need for more empirical evaluation to fully assess the viability of the

assumptions involved and the accuracy of the two prophecy formulas in the IRT and CTT contexts.

Appendix

In this appendix, we offer a derivation of Formula 15 from the classical test theory (CTT) perspective. Let the reliability of test $x(\rho_{xx})$ with n_x items be written as

$$\rho_{xx} = \frac{\sigma_x^2 - SEM_x^2}{\sigma_x^2}, \tag{A1}$$

where SEM_x^2 represents the (group-level) variance of measurement error. According to Lord and Novick (1968), the group-level variance of measurement error can be expressed as the average of individual-level variance of measurement error. That is,

$$SEM_x^2 = E\left(SEM_{sx}^2 \right), \tag{A2}$$

where s stands for an individual or subject and where the expectation (E) is taken over all subjects in a group or population. Because x is the sum of n_x item scores and because the item-level measurement errors are assumed uncorrelated with each other within the CTT framework (Lord & Novick, 1968), we can write x and SEM^2 for a given subject (s) as follows:

$$x_s = x_{s1} + \ldots + x_{sn_x} \tag{A3}$$

$$SEM_{sx}^2 = SEM_{s1}^2 + \ldots + SEM_{sn_x}^2 = \sum_{i=1}^{n_x} SEM_{si}^2 . \tag{A4}$$

Instead of the total score, let us now consider the proportional score, which can be written simply as

$$p\left(x_s\right) = \frac{x_s}{n_x} = \frac{x_{s1}}{n_x} + \ldots + \frac{x_{sn_x}}{n_x}. \tag{A5}$$

Because $p(x_s)$ is a simple linear transformation of x_S, the reliability of a proportional score will be the same as that of a total score. That is, in view of Equations A1 through A5,

$$\rho_{xx} = \rho_{\rho(x)\rho(x)} = \frac{\frac{\sigma_x^2}{n_x^2} - E\left(\sum_{i=1}^{n_x} \frac{SEM_{si}^2}{n_x^2} \right)}{\frac{\sigma_x^2}{n_x^2}} = \frac{\frac{\sigma_x^2}{n_x^2} - \left(\frac{1}{n_x} \right) E\left(\sum_{i=1}^{n_x} \frac{SEM_{si}^2}{n_x} \right)}{\frac{\sigma_x^2}{n_x^2}} . \tag{A6}$$

Again, the expectation is taken over all subjects in a group. Let us now assume that there is another test, y, with n_y items, which is obtained by adding items to, or deleting items from, x. As previously stated, the added or deleted items are assumed to be similar in content and psychometric quality to the items already in test x. In view of Equation A6, the reliability of proportional scores on $y[p(y)]$ can be expressed as

$$\rho_{yy} = \rho_{\rho(y)\rho(y)} = \frac{\frac{\sigma_y^2}{n_y^2} - E\left(\sum_{i=1}^{n_y} \frac{SEM_{si}^2}{n_y^2}\right)}{\frac{\sigma_y^2}{n_y^2}} = \frac{\frac{\sigma_y^2}{n_y^2} - \left(\frac{1}{n_y}\right) E\left(\sum_{i=1}^{n_y} \frac{SEM_{si}^2}{n_y}\right)}{\frac{\sigma_y^2}{n_y^2}}. \quad \text{(A7)}$$

In Equation A6, the average measurement error variance of items can be expressed as:

$$Ave\left(SEM_{i_x}^2\right) = E\left(\sum_{i=1}^{n_x} \frac{SEM_{si}^2}{n_x}\right), \quad \text{(A8)}$$

where the notation i_x on the left hand side is used to emphasize the fact that item i is from test x. In view of Equation A8, Equation A6 can be rewritten as

$$\rho_{xx} = \rho_{\rho(x)\rho(x)} = \frac{\frac{\sigma_x^2}{n_x^2} - \left(\frac{1}{n_x}\right)\left(Ave\left[SEM_{i_x}^2\right]\right)}{\frac{\sigma_x^2}{n_x^2}}. \quad \text{(A9)}$$

Solving for the average item-level measurement error variance in x, one obtains

$$Ave\left(SEM_{i_x}^2\right) = n_x\left(\frac{\sigma_x^2}{n_x^2}\right)\left(1-\rho_{xx}\right). \quad \text{(A10)}$$

Now turning to test y, Equation A7 can be rewritten as

$$\rho_{yy} = \rho_{\rho(y)\rho(y)} = \frac{\frac{\sigma_y^2}{n_y^2} - \left(\frac{1}{n_y}\right)\left(Ave\left[SEM_{i_y}^2\right]\right)}{\frac{\sigma_y^2}{n_y^2}}. \quad \text{(A11)}$$

Solving for the average item-level measurement error variance in y, one obtains

$$Ave\left(SEM_{i_y}^2\right) = n_y\left(\frac{\sigma_y^2}{n_y^2}\right)\left(1-\rho_{yy}\right). \quad \text{(A12)}$$

If we assume that average item-level SEM^2s are equal in tests x and y, one obtains, in view of Equations A10 and A12,

$$n_x\left(\frac{\sigma_x^2}{n_x^2}\right)\left(1-\rho_{xx}\right) = n_y\left(\frac{\sigma_y^2}{n_y^2}\right)\left(1-\rho_{yy}\right). \quad \text{(A13)}$$

It should be noted that this assumption (the average item-level SEM^2s from x and y are equal) is very similar to the assumption made within the item response theory (IRT) context: The average item information functions are equal. Solving for ρ_{yy},

$$\rho_{yy} = \frac{\rho_{xx} + kA - 1}{kA}, \tag{A14}$$

where

$$k = \frac{n_y}{n_x} \tag{A15}$$

and

$$A = \frac{\left(\sigma_y^2\right) / n_y^2}{\left(\sigma_x^2\right) / n_x^2}. \tag{A16}$$

The numerator and denominator in Equation A16 above refer to the variances of proportional scores in y and x, respectively. Because items in x and y are, by definition, very comparable, it is not unreasonable to assume that the variances of proportional scores are equal; that is, $A = 1$. With this assumption, Equation A14 can be rewritten as

$$\rho_{yy} = \frac{\rho_{xx} + \left(k - 1\right)}{k} \tag{A17}$$

which is identical to Equation 15. The assumption that $A = 1$ is similar to the assumption previously used within the IRT framework (see Equation 14). The data (standard deviations of proportional scores) in the last column of Table 1 offer some empirical support for this assumption ($A = 1$) in the current example. There is still a need for a comprehensive empirical verification of this assumption. Equation 15 is derived within the IRT framework, whereas the current equation (Equation A17) is derived within the CTT framework. The two underlying assumptions are also very similar in the two frameworks.

References

Divgi, D. R. (1989). Estimating reliabilities of computerized adaptive tests. *Applied Psychological Measurement, 13*, 145-149.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*, 233-245.

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

Mislevy, R. J., & Bock, R. D. (1990). *BILOG3: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement, 23*, 239-247.

Samejima, F. (1977). A use of information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229-244.

Sympson, J. B. (1980, April). *Estimating the reliability of adaptive tests from a single administration*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa.