Measuring the Power of Learning.®

Research Report
ETS RR–17-02

Accuracy of a Classical Test Theory–Based Procedure for Estimating the Reliability of a Multistage Test

Sooyeon Kim

Samuel A. Livingston

December 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Accuracy of a Classical Test Theory–Based Procedure for Estimating the Reliability of a Multistage Test

Sooyeon Kim & Samuel A. Livingston

Educational Testing Service, Princeton, NJ

The purpose of this simulation study was to assess the accuracy of a classical test theory (CTT)–based procedure for estimating the alternate-forms reliability of scores on a multistage test (MST) having 3 stages. We generated item difficulty and discrimination parameters for 10 parallel, nonoverlapping forms of the complete 3-stage test and ability parameters for a population of 30,000 simulated test takers. Using these parameters to generate item responses, we ran each of the 30,000 simulated test takers through each of the 10 simulated forms of the 3-stage test and computed the correlation of the scores on each pair of simulated test forms. We then computed the CTT estimate of the reliability, in the full population of 30,000 simulated test takers, of the total scaled scores resulting from the multistage testing procedure. We computed the estimate separately from the simulated responses to each of the 10 simulated forms of the 3-stage test. The reliability estimates from each simulated form of the MST differed by less than .005 from the average of the correlations between scores on that form and on the other 9 simulated forms of the MST.

Keywords Multistage test; reliability; classical test theory; item response theory; simulation

doi:10.1002/ets2.12129

Multistage testing is an adaptive testing procedure in which the test is divided into two or more stages. The items presented to the test taker at each stage (except the first) depend on the test taker's performance on the previous stages. Multistage testing differs from the procedure commonly called computer-adaptive testing (CAT) by having only a small number of decision points—in some cases, only one—whereas CAT has a decision point after each item.

CAT item selection algorithms construct each test form while the test taker is taking the test by iteratively administering an item, estimating a provisional score, and then selecting the next item from the active item bank using certain statistical optimization criteria (Luecht & Nungester, 1998). Under multistage testing, however, there is a predefined grouping of items into modules based on content and statistical specifications. Multistage test (MST) construction enables the test developers and test form assemblers to carefully scrutinize all the modules and test forms to achieve desired test characteristics (e.g., distribution of item content and difficulty).

Figure 1 displays an example of a three-stage MST form in which two adaptations to the test takers' ability levels take place. At Stage 1 (often called routing), there is only one module; all test takers taking that form of the test are tested with same set of items. At Stage 2, there are two modules: a high-difficulty module and a low-difficulty module. The items a test taker receives at Stage 2 are determined by the test taker's performance on Stage 1. At Stage 3, there are three modules: a high-difficulty module, a medium-difficulty module, and a low-difficulty module. The items a test taker receives at Stage 3 are determined by the test taker's performance on Stages 1 and 2.

We will use the term *variant* to mean a combination of modules that could possibly be presented to a test taker. In the example of Figure 1, each variant consists of the first-stage module, one of the second-stage modules, and one of the third-stage modules. There are four variants of each form of the test illustrated in Figure 1:

Variant 1: Module 1 (routing), Module 2D (difficult), Module 3D (difficult).
Variant 2: Module 1 (routing), Module 2D (difficult), Module 3M (medium difficulty).
Variant 3: Module 1 (routing), Module 2E (easy), Module 3M (medium difficulty).
Variant 4: Module 1 (routing), Module 2E (easy), Module 3E (easy).

Corresponding author: S. Kim, E-mail: skim@ets.org

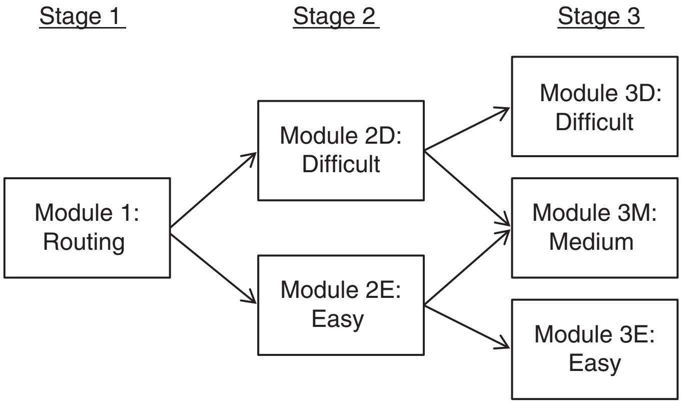

Figure 1 Schematic of a three-stage multistage test.

Reliability of What Score?

There is more than one way to score an MST. The simplest way is to use number-correct scoring. The reliability estimation procedure evaluated in this study is intended for MSTs that are scored by counting the number of correct answers.[1] In the MST illustrated in Figure 1, the raw score on any of the four variants is simply the total number of items answered correctly on the three stages. In each form of the test, each variant has its own raw-to-scale score conversion, determined in a way that makes the scores on any two variants comparable, in the portion of the score range where a test taker might be tracked to either variant.

Reliability in What Population?

The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education's *Standards for Educational and Psychological Testing* (2014) defined reliability as "the consistency of scores across replications of testing procedure" (p. 33). In the case of an MST, the appropriate population is the population of test takers who are to be tested with that multistage testing procedure. The complication in estimating the alternate-forms reliability of an MST arises from the possibility that a test taker might be tracked to different modules on two replications of the multistage testing procedure. For example, a test taker might be tracked to the high-difficulty third-stage module on one testing and to the medium-difficulty third-stage module on another testing. However, the score users to whom the scores are reported do not see the test taker's score on each individual module. They see a single scaled score, reported on a scale that does not depend on the variant of the test—the particular combination of modules—that the test taker took. They do not know what variant of the test (what combination of modules) each test taker took, and they do not need to know, because the scores on the different variants of the MST are scaled for comparability. The usual population for which to estimate the reliability of the reported scores is the population of all test takers taking the test. In the case of an MST, that population includes all test takers, regardless of what variant of the test each test taker took.

The alternate-forms reliability coefficient of a test in a population of test takers is the correlation of the scores that would result if each test taker in the population were to take two forms of the test, with no practice effect and no change in any test taker's true ability. If the test is an MST, the alternate-forms reliability coefficient is the correlation of the scores that would result if each test taker in the population were to take two different forms of the complete MST, following the specified multistage procedure each time. A form of the MST consists of a specific set of items for each module and a set of tracking rules (cutscores) for progressing from one module to the next. The problem is to estimate the alternate-forms reliability of the scaled scores resulting from the multistage testing procedure, in the full population of test takers for the MST.

Relevant Previous Work

Green, Bock, Humphreys, Linn, and Reckase (1984) expressed the opinion that classical test theory (CTT) is not suited to adaptive tests, because the classical reliability index is relevant when all test takers take the same set of test items. In their view, the definition of reliability has little relevance for measurement based on item response theory (IRT), where the error variance is expressed as a function of ability. In general, IRT scoring emphasizes conditional measurement precision, as shown by the test information function, rather than average indices across the score scale, such as reliability. Because reliability has been widely used in practice as an established criterion for test quality, however, other articles have dealt with reliability in a model-based IRT context (Samejima, 1994) as well as in the adaptive testing context (Nicewander & Thomasson, 1999; Zhang, Breithaupt, Tessema, & Chuah, 2006) and have compared measurement precision under the CTT and IRT frameworks (Mellenbergh, 1996; Thissen, 2000). Particularly, Zhang et al. (2006) compared two IRT-based procedures that can be used to estimate test reliability for both adaptive (MST) and nonadaptive testing designs using a certification exam. Many adaptive testing programs use a model-based IRT approach to estimate overall reliability coefficients (for more detailed information, see van Rijn, 2014). However, we did not find any proposed solutions for estimating MST reliability based on CTT in the psychometric literature, including the reliability chapter of the most recent edition of *Educational Measurement* (Haertel, 2006).

Recently, Livingston and Kim (2014) introduced the MST reliability estimation procedure evaluated in the present study. The procedure estimates the reliability of scaled scores that are computed from the number-correct raw scores, using a different raw-to-scale conversion for each variant (i.e., each possible path) of a two-stage MST. The reliability estimation procedure is based on two assumptions. One assumption is that the reliability of scores on each test module, in the group of test takers who take that module, can be estimated accurately. (This group is not the population of test takers who might possibly have taken that module but the group of test takers who actually took it.) The other assumption is that the linking of the scores on different variants of the MST is accurate.^{2} If the raw-to-scale conversion for each variant of each form of the test is accurate, the scaled scores on all the variants of a test form will be comparable to each other (and to scores on other forms of the test).

Livingston and Kim (2014) assessed the accuracy of their estimation procedure in a simulation study based on an MST modeled after the *GRE*® revised General Test in Verbal Reasoning and Quantitative Reasoning. Each of those tests is a two-stage test with three modules in the second stage, so that there are three variants of each form of the test. To create a situation in which the correct value of the reliability coefficient could be observed directly, they generated data for a group of simulated test takers taking two forms of the MST. In their simulation, the single-form reliability estimates from their estimation procedure differed by less than .005 from the correlations of the scores on two forms of the MST.

The Present Study

The purpose of the present study was to assess the accuracy of the CTT reliability estimation procedure for a three-stage test. The test used in this study followed the format of the test illustrated in Figure 1, with 15 items in each module. Following the simulation procedure of the previous study by Livingston and Kim (2014), we created a situation in which a group of simulated test takers took 10 separate forms of the MST. For each of the 10 forms of the MST, we computed a reliability estimate based on the data from only that form. We compared that reliability estimate with the correlations between scores on that form and on the other nine forms of the MST. We also compared the average of the 10 reliability estimates with the average of the 45 possible correlations between different forms of the MST.

The MST reliability estimation procedure proposed by Livingston and Kim (2014) resembles the derivation of the composite reliability formula (e.g., Haertel, 2006, p. 76), in that it estimates the variance of errors of measurement (VEM) separately for different parts of the full test and then combines those estimates. In the MST case, however, the parts of the full MST are taken by different groups of test takers.

To describe the reliability estimation procedure for the three-stage test shown in Figure 1, we will refer to the groups of test takers taking the four variants of the MST as Group 1, Group 2, and so on. (These are the groups of test takers actually taking the four variants on the form of the MST for which data are available. They are not the groups of test takers who might take the corresponding variants on some other form of the MST.) In our notation, we will use the subscript $i = 1, 2, 3$, or 4 to identify these groups of test takers. The subscript *all* will represent the combined group of all test takers taking the MST. We will use n for the number of test takers, "rel" for reliability coefficient, and "var" for variance. "SEM" will represent the standard error of measurement; "VEM" will represent the VEM. The abbreviations "raw" and "scale" will indicate whether a statistic applies to raw scores or scaled scores.

The reliability of the scaled scores in the full group of test takers is

$$rel_{\text{all}}\,(\text{scale}) = 1 - \frac{VEM_{\text{all}}\,(\text{scale})}{var_{\text{all}}\,(\text{scale})}. \tag{1}$$

The error of measurement in a test taker's score is independent of the error of measurement in any other test taker's score. The VEM for a group of test takers is the average of the conditional VEMs for the individual test takers. Therefore the VEM of the scaled scores for the combined group is the weighted average of the VEMs in the four separate groups, weighting the VEM for each group by the number of test takers in the group:

$$VEM_{\text{all}}\,(\text{scale}) = \frac{\sum_{i=1}^{4} n_i VEM_i\,(\text{scale})}{\sum_{i=1}^{4} n_i}. \tag{2}$$

We will need an estimate of the VEM of the scaled scores in each of the four groups. The estimation procedure is the same for each group. To estimate the SEM of the scaled scores of Group *i*, multiply the SEM of their raw scores by the slope of the raw-to-scale conversion for the variant of the test taken by Group *i*. If the conversion is not linear, the slope will not be constant, but on the average, it will be approximately equal to the ratio of the standard deviations:

$$SEM_i\,(\text{scale}) \cong SEM_i\,(\text{raw})\,\frac{SD_i\,(\text{scale})}{SD_i\,(\text{raw})}. \tag{3}$$

Squaring both sides of the equation,

$$VEM_i\,(\text{scale}) \cong VEM_i\,(\text{raw})\,\frac{var_i\,(\text{scale})}{var_i\,(\text{raw})}. \tag{4}$$

To estimate the VEM of the raw scores of Group *i*, estimate the reliability of the raw scores by any appropriate method. Then estimate the VEM by

$$VEM_i\,(\text{raw}) = var_i\,(\text{raw})\left[1 - \text{rel}_i\,(\text{raw})\right]. \tag{5}$$

Substituting that estimate into Equation 4,

$$VEM_i\,(\text{scale}) \cong \left[1 - rel_i\,(\text{raw})\right] var_i\,(\text{scale}). \tag{6}$$

This procedure provides an estimate of the VEM of scaled scores in the group of test takers taking each variant of the MST. Inserting these estimates into Equation 2 provides an estimate of the VEM of scaled scores in the full group of all test takers. Inserting that estimate into Equation 1, we have an estimate of the reliability coefficient of the scaled scores in the group of all test takers.

Testing the Accuracy of the Estimation Procedure

We created 30,000 simulated test takers by generating their ability parameters from the standard normal distribution (i.e., $\theta \sim N\ (0,\ 1)$). Because we intended to use the two-parameter logistic (2PL) IRT model to generate the responses of the simulated test takers to the simulated items, we generated a discrimination parameter (a) and a difficulty parameter (b) for each simulated item.[3] We used the computer program WinGen (Han, 2007) to generate test taker ability parameters and item difficulty and discrimination parameters. WinGen allows the user to specify, for each of the two parameters, the mean and standard deviation of the pool from which values are sampled. To make our simulation as realistic as possible, we specified values comparable to those from an actual operational test.

We generated 10 test forms one at time, using the same statistical specifications for each form, to make them as parallel as possible. Each of the six modules in a particular form was also created separately, one after another, based on its own statistical specifications. With only 15 items per module, however, the difficulty and discrimination of the items in a module differed slightly across the 10 forms. Table 1 presents the range, over the 10 simulated forms, of means and standard deviations of the item discrimination parameters (a) and the item difficulty parameters (b) of the items in each of the

Table 1 Range of Means and Standard Deviations of Item Discrimination and Item Difficulty for Each of the Six Modules Over the 10 Simulated Forms

		a Parameter (discrimination)		*b* Parameter (difficulty)	
Stage	Class/level	*M* (min, max)	*SD* (min, max)	*M* (min, max)	*SD* (min, max)
1	Routing	(0.75, 0.90)	(0.21, 0.38)	(−0.03, 0.02)	(0.35, 0.90)
2	Difficult	(0.78, 0.99)	(0.18, 0.41)	(0.51, 0.61)	(0.40, 0.67)
2	Easy	(0.75, 0.96)	(0.19, 0.36)	(−0.59, −0.51)	(0.51, 0.74)
3	Difficult	(0.87, 1.15)	(0.22, 0.35)	(1.00, 1.12)	(0.62, 0.84)
3	Medium	(0.72, 0.88)	(0.13, 0.26)	(−0.09, 0.12)	(0.62, 1.06)
3	Easy	(0.88, 1.07)	(0.27, 0.38)	(−1.14, −1.01)	(0.45, 0.91)

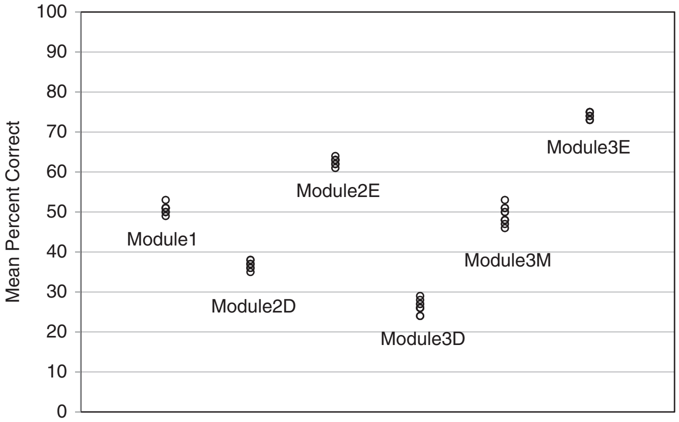

Figure 2 Mean percent correct sores of all simulated test takers for each of the six modules.

six modules. Figure 2 displays, separately for each module, a comparison of the difficulty of the 10 forms based on the data from the simulation. For each module, the figure shows 10 data points, one for each of the 10 forms. The difficulty differences among the forms were small, except in the medium-difficulty Stage 3 module. We used SAS statistical software to generate each simulated test taker's dichotomous response (correct or incorrect) to each item of a particular module.

Each form of the MST consisted of a Stage 1 module, two Stage 2 modules, three Stage 3 modules, three cutscores, and four raw-to-scale conversions. The simulation procedure for each test taker, for each of the 10 MST forms, consisted of the following six steps:

1. Generate the simulated test taker's response to each item in the Stage 1 module.
2. Compute the test taker's number-correct raw score on Stage 1 and apply the Stage 1 cutscore to assign the test taker to the appropriate Stage 2 module.
3. Generate the test taker's responses to each item in the Stage 2 module, and compute the test taker's number-correct raw score.
4. Compute the test taker's number-correct raw score on Stages 1 and 2 combined and apply the appropriate Stage 2 cutscore to assign the test taker to the appropriate Stage 3 module.
5. Compute the test taker's number-correct raw scores on Stage 1, Stage 2, and Stage 3 and the total raw score.
6. Apply the appropriate raw-to-scale conversion to the test taker's total raw score to determine the test taker's scaled score.

We replicated this procedure for 30,000 test takers, each taking 10 forms of the MST.

Each form of the MST had four raw-to-scale conversions to convert the number-correct raw scores on the four variants of that form to scaled scores on a scale from 100 to 200. The raw-to-scale conversions were determined by IRT true-score equating using the 2PL model. Figure 3 shows the test characteristic curve (TCC) for each of the four variants of one of the 10 forms of the MST. Figure 4 shows the raw-to-scale score conversions for those four variants of this same form. Although the conversions appear strongly curvilinear, most of the nonlinearity is in the portions of the score range where there are

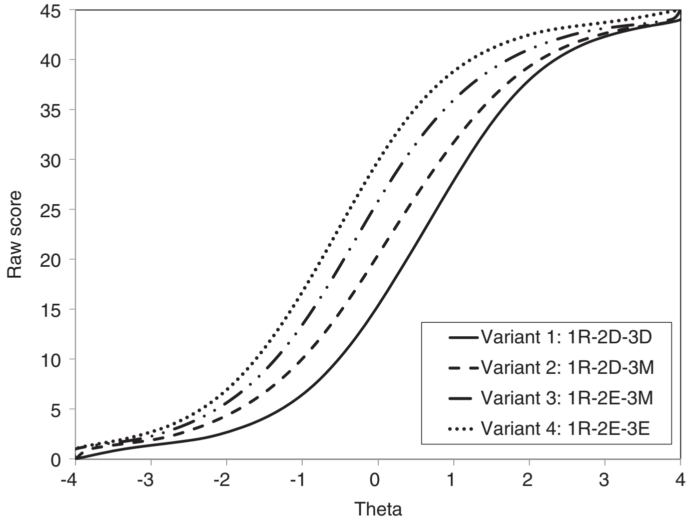

Figure 3 Test characteristic curves for each of the four variants on Form 1.

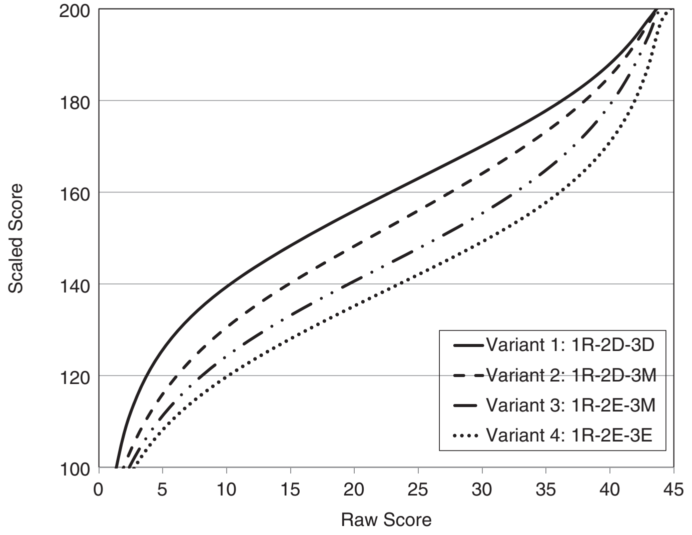

Figure 4 Scaled score conversions for each of the four variants on Form 1.

very few test takers' scores. The conversion for each module is approximately linear in the portion of the range containing the raw scores of most of the test takers who actually take that module. The pattern of the TCCs was very similar across the forms. We selected the cutscores for routing test takers to the different modules so as to result in approximately one fourth of the test takers taking each variant. On average, the proportions of test takers taking Variant 1, 2, 3, or 4 were 25%, 26%, 24%, and 25%, respectively, with very little variation in these percentages over the 10 forms of the MST.

We applied the CTT reliability estimation procedure separately to the item response data from each of the 10 forms. The procedure requires, as input, an estimate of the reliability (or the SEM or the VEM) of scores on each of the four variants of the test, in the group of test takers taking that variant. (See Equations 4 and 5.) For each of these estimates, we computed coefficient alpha[4] separately for each module and used it to estimate the VEM on that module in a group of test takers taking that variant of the MST. Because errors of measurement on any module are independent of errors of measurement on any other module, the VEM of the total raw score on any variant of the MST is the sum of the VEMs on the three modules in that variant:

$$VEM_i\,(\text{raw}) = VEM_i\left(\text{stage }1\right) + VEM_i\left(\text{stage }2\right) + VEM_i\left(\text{stage }3\right). \tag{7}$$

This estimate of the VEM of the raw total scores could then be entered into Equation 4.

Applying this reliability estimation procedure separately to the data for each of the 10 forms of the MST gave us 10 independent estimates of the reliability coefficient of the MST in the group of all 30,000 simulated test takers. For each form of the MST, we compared this estimated reliability coefficient with the correlations of that form of the MST with the other nine forms.

For the sake of comparison, we used an IRT procedure to estimate the reliability coefficient of the same 10 MST forms. The IRT reliability estimates for the scaled scores were obtained using the same 30,000 simulated test takers according to the following formula:

$$rel_{\text{all}}\,(\text{scale}) = 1 - \frac{VEM_{\text{all}}\,(\text{scale})}{var_{\text{all}}\,(\text{scale})} = 1 - \frac{\sum_{\theta=-3.0}^{+3.0} w_\theta\left(\text{CSEM}_\theta\right)^2}{var_{\text{all}}\,(\text{scale})} = 1 - \frac{\sum_{\theta=-3.0}^{+3.0} w_\theta\left(\text{SD}_\theta\right)^2}{var_{\text{all}}\,(\text{scale})} \tag{8}$$

where θ indexes the intervals in a partition of the true ability scale of -3.0 to $+3.0$ into intervals of size 0.1. CSEM_θ indicates the conditional SEM estimated at the midpoint of the interval. The weight w_θ is the proportion of the simulated test takers whose ability values were in the interval.

How Accurate Were the Estimates?

Table 2 shows the comparison of the actual alternate-form correlations with the CTT reliability estimates, along with the IRT reliability estimates. For each of the 10 forms of the MST, the difference between the estimated reliability and the mean of the correlations with the other forms is very small—no larger than .004. And for each of the 10 forms, the CTT reliability estimate is slightly higher than the mean of the actual correlations. Notice in Table 2 that the correlations and reliability estimates are so consistent from one test form to another that it is necessary to use three decimal places to show any differences. If the numbers were reported to two decimal places, as is the usual practice, all the correlations and nearly all the reliability estimates would be .94. Although the systematic overestimation of reliability by the CTT procedure was so small as to be negligible for practical purposes, it could be worthwhile to determine the reason for the overestimation. The appendix presents some intermediate results produced at selected stages of the CTT reliability estimation procedure.

Discussion

The reliability estimation procedure we evaluated in this study is an attempt to use CTT to estimate the correlation of the scores on two complete replications of the multistage testing procedure in the full test taker population. Any attempt to estimate the reliability of a single variant of the test or a single module immediately raises the question of how the test taker population is to be defined. Some test takers will be tracked to different variants and to different modules on two replications of the MST. These test takers would be a nonnegligible proportion of the people who take the MST unless

Table 2 Estimated and Actual Reliability Coefficients

MST form	Range of correlations with other forms	Mean of correlations with other forms	CTT reliability estimate	IRT reliability estimate
1	.937–.941	.939	.941	.938
2	.938–.942	.940	.944	.941
3	.936–.940	.938	.938	.936
4	.939–.943	.941	.944	.942
5	.938–.943	.940	.944	.941
6	.938–.942	.940	.943	.940
7	.940–.943	.941	.945	.943
8	.937–.941	.939	.942	.938
9	.936–.941	.939	.942	.938
10	.936–.940	.938	.941	.937
All 10 forms	.936–.943	.940	.942	.939

Note. CTT = classical test theory; IRT = item response theory; MST = multistage test.

routing modules at any stage include plenty of highly discriminating items. We do not think it makes sense to define the test taker population in a way that includes a test taker on one replication of the testing procedure and excludes that same test taker on another replication of the testing procedure. They must be included in estimating the reliability of the MST in the population of test takers who take it.

The purpose of the present study was to assess the accuracy of a procedure based on CTT for estimating the alternate-forms reliability of scores on a three-stage MST. The CTT method produced highly accurate estimates. The present findings were consistent with those reported by Livingston and Kim (2014) for a two-stage MST. The present findings provide additional evidence for the generalizability of the CTT estimation method and may justify its operational use.

For the sake of comparison, we used an IRT procedure to estimate the reliability of the same 10 MST forms. Like the CTT procedure evaluated in this study, the IRT procedure yielded accurate reliability estimation results for the same 10 three-stage MST forms. (The IRT procedure actually yielded slightly more accurate reliability estimation results than did the CTT procedure — for these simulated responses, which were generated by that same IRT model.) For an MST on which all the items have been calibrated with an appropriate IRT model, the IRT and CTT approaches to reliability estimation may work equally well. The advantage of the CTT procedure over IRT procedures for reliability estimation is that it does not require strong assumptions or relatively large datasets. It can also be computed more easily and explained more simply than IRT procedures. The reliability coefficient can be estimated for an MST with number-correct scoring, as it can for a conventional nonadaptive test. The present study shows that the estimate is likely to be highly accurate.

Notes

1 The procedure was developed for use with the GRE revised General Test Verbal Reasoning and Quantitative Reasoning subtests.

2 A scoring procedure that fails to meet this assumption would create problems far more serious than inaccurate estimates of reliability.

3 The 2PL model is used operationally for some large-scale testing programs (e.g., the GRE, *TOEFL®*, and *TOEIC®* tests).

4 Coefficient alpha was an appropriate estimate of alternate-forms reliability for our simulated data. It would not be an appropriate estimate if each module were to consist of two or more item sets (i.e., groups of items based on common stimulus material).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65 – 110). Westport, CT: Praeger.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 20,* 369 – 377.

Livingston, S. A., & Kim, S. (2014). Multistage test reliability estimated via classical test theory. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 265 – 270). Boca Raton, FL: CRC Press.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229 – 249.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293 – 299.

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

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

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159 – 184). Hillsdale, NJ: Lawrence Erlbaum.

van Rijn, P. (2014). Reliability of multistage tests using item response theory. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 251 – 264). Boca Raton, FL: CRC Press.

Zhang, Y., Breithaupt, K., Tessema, A., & Chuah, D. (2006, April). *Empirical vs. expected IRT-based reliability estimation in computerized multistage testing.* Paper presented at the annual conference of the National Council of Measurement in Education, San Francisco, CA.

Appendix

Table A1. Range of Descriptive Statistics for Each of the Four Variants Over the 10 Simulated Forms

		Raw score				Scaled score		
Variant	*n*	*M*	*SD*	Reliability	SEM	*M*	*SD*	SEM
1. Routing-difficult-difficult	7,041 – 7,931	30.5 – 31.2	5.0 – 5.4	0.70 – 0.74	2.7 – 2.8	171.2 – 172.4	8.5 – 9.0	4.3 – 4.9
2. Routing-difficult-medium	7,449 – 8,708	22.5 – 23.7	3.6 – 4.3	0.43 – 0.59	2.6 – 2.7	152.3 – 154.3	5.3 – 6.1	3.8 – 4.3
3. Routing-easy-medium	6,559 – 8,033	20.5 – 22.1	3.6 – 4.0	0.46 – 0.56	2.6 – 2.7	142.4 – 144.0	5.2 – 6.1	3.7 – 4.3
4. Routing-easy-easy	6,861 – 7,927	13.4 – 14.7	4.8 – 5.3	0.69 – 0.75	2.7 – 2.8	124.7 – 126.2	8.9 – 9.2	4.6 – 5.1

Suggested citation:

Kim, S., & Livingston, S. A. (2017). *Accuracy of a classical test theory – based procedure for estimating the reliability of a multistage test* (ETS Research Report No. RR-17-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12129

Action Editor: Rebecca Zwick

Reviewers: Neil Dorans and Tim Moses