Research Report

Estimating Reliability for Tests With One Constructed-Response Item in a Section

ETS RR–24-07

Yanxuan Qu
Sandip Sinharay

December 2024

ETS Research Memorandum 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.

ETS Research Report Series ISSN 2330-8516

ETS RESEARCH REPORT

Estimating Reliability for Tests With One Constructed-Response Item in a Section

Yanxuan Qu[1] & Sandip Sinharay[2]

1 ETS Psychometric Analysis and Research, ETS, Princeton, New Jersey, United States
2 ETS Research Institute, ETS Princeton, New Jersey, United States

The goal of this paper is to find better ways to estimate the internal consistency reliability of scores on tests with a specific type of design that are often encountered in practice: tests with constructed-response items clustered into sections that are not parallel or tau-equivalent, and one of the sections has only one item. To estimate the reliability of scores on this kind of test, we propose a two-step approach (denoted as CA_STR) that first estimates the reliability of scores on the section with a single item using the correction for attenuation method and then estimates the reliability of scores on the whole test using the stratified coefficient alpha. We compared the CA_STR method with three other reliability estimation approaches under various conditions using both real and simulated data. We found that overall, the CA_STR method performed the best and it was easy to implement.

Keywords constructed-response items; test reliability; correction for attenuation; stratified coefficient alpha; single-item reliability

doi:10.1002/ets2.12382

The topic of reliability has interested measurement practitioners arguably since Spearman (1910) introduced the concept of correction for attenuation (CA) and that of split half reliability. Reliability is traditionally quantified using reliability coefficients. One class of these coefficients includes internal consistency estimates, which are computed from a single administration. Most internal-consistency reliability estimates, such as split half, KR-20, and Cronbach's alpha, assume that the test items (or sections or components) are essentially tau-equivalent (e.g., Allen & Yen, 2002, p. 91; Feldt & Qualls, 1996; Novick & Lewis, 1967), which means that the true scores on any pair of items are perfectly correlated and differ by the same constant for all examinees. However, educational and psychological tests often have multiple sections that vary in length, item type, task complexity, and score ranges. Some tests may have only one item in a section. An example of the structure of such a test is shown in Table 1. The items in any section are of different types than those in any other section, and Section 1 of the test includes only one item whose score range (0–2) is different from the score range of all items on the test. While it is necessary to report reliability of all reported scores, it is not clear how to estimate internal consistency reliability for these tests that (a) have nonparallel sections and (b) have single-item sections such that the single item is of a different type than all other items on the test.

The purpose of this study is to find a practical way to better estimate the reliability of scores on a constructed-response (CR) test that includes multiple sections that are not parallel in length or content and one of the sections has only one item that is of a different type than the other items on the test.

Internal Consistency Reliability Estimates: A Literature Review

Researchers have developed several approaches to estimate the internal consistency reliability of total test scores under different test designs. Kuder and Richardson (1937) suggested Kuder-Richardson Formula 20 (described in Appendix A) for estimating internal consistency reliability for tests with only dichotomous items. Guttman (1945) proposed six Lambdas (Appendix A) to calculate lower bounds of reliability for tests with dichotomous or polytomous items. Among the six estimates, λ_2 and λ_3 are the most convenient to use and are both larger than λ_1. Cronbach (1951; see Formula 4 later in this paper) introduced the alpha statistic for estimating internal consistency reliability for tests with either dichotomous or polytomous items. Cronbach's alpha is very similar to Guttman's λ_3 or λ_2. All of these approaches require items on a test to be essentially tau-equivalent and homogeneous with respect to content (e.g., Allen & Yen, 2002, p. 91; Novick &

Corresponding author: Yanxuan Qu, Email: yqu@ets.org

Table 1 An Example of the Structure of a Constructed-Response (CR) Test With Nonparallel Sections and a Single-Item Section

Section	Number of items	Item score range	Item sequence
Section 1	1	0–2	1
Section 2	3	0–3	2–4
Section 3	2	0–4	5–6

Lewis, 1967). These approaches will provide inaccurate and underestimated reliability coefficients when the items are not tau-equivalent or not content-homogeneous, which may occur when, for example, items have different score categories and belong to different content areas. Cronbach et al. (1965) introduced the stratified alpha coefficient (see Formula 3 of this paper), which was proved to be more accurate and robust than Cronbach's alpha when test content is not homogeneous (Feldt & Qualls, 1996). Kristof (1974) proposed an approach to estimate score reliability for tests with three non-parallel or non-tau-equivalent parts in length (Formulas 5 and 6 of this paper). Using real data, Kristof (1974) illustrated that the Kristof reliability estimate was at least as accurate as Cronbach's alpha and Guttman's lower bound λ_3. Sedere and Feldt (1977) used simulated item scores on a test with three sections of different lengths to evaluate the accuracy of the Kristof reliability coefficient compared to those of Cronbach's alpha and Guttman's λ_2. They found that the Kristof's coefficient can be sensitive to part-test length ratios or heterogeneity of the three sections in length but can still outperform Cronbach's alpha and Guttman's λ_2 when the three sections are moderately heterogeneous in length and sample size is at least 200. The Kristof method also requires content homogeneity within a test. Sedere and Feldt (1977) did not consider the effect of test dimensionality.

Molenaar and Sijtsma (1988) suggested a reliability estimate, which will be referred to as the MS estimate, for both dichotomous and polytomous data. Their approach involved less restrictive assumptions compared to those underlying Cronbach's alpha or Guttman's λ_2, but required unidimensionality and double monotonicity[1] (Sijtsma & Molenaar, 2002), which are still pretty strong assumptions. van der Ark et al. (2011) proposed the latent class reliability coefficient (LCRC) and compared it with four reliability estimates—Cronbach's alpha, Guttman's λ_2, the MS estimate, and the split-half reliability coefficient—using data simulated from unidimensional and multidimensional graded response models. They varied test length, item score format (dichotomous or polytomous), discrimination parameters (equal or unequal), and sample size. They found that the MS estimate and Guttman's λ_2 had the least bias for unidimensional tests with equal discrimination parameters. But for multidimensional data or data with unequal discrimination parameters, the LCRC was less biased than the others. Estimation bias was smaller for all methods when the test was unidimensional or had more items.

Cronbach et al. (1972) introduced the generalizability theory (G-theory), a different framework for estimating reliability of test scores. The G-theory uses the analysis of variance (ANOVA) technique to break down the measurement error in classical test theory into multiple error sources from different testing situations (e.g., raters, times, items). The G-theory can estimate various types of test reliabilities, internal consistency reliability being one of them, but it does not address the issue of having only one item in a test section. Brennan (2017) conceptually explained why it can be problematic when using the G-theory to estimate reliability for a test with only one item in a section. According to Brennan (2017, p. 3), the generalizability study based on a single condition of a facet leads to bias in error variances and coefficients. The fundamental cause of this bias is that the single fixed level of a facet in a generalizability study makes it impossible to disentangle some variance components.

All the aforementioned methods were initially developed to estimate reliability at the test level. Over the years, researchers have devoted their efforts not only to enhancing the estimation of test-score reliability under various situations, but also to exploring approaches for estimating the reliability of a single-item score/measure. Wanous and Reichers (1996) proposed a method for estimating the reliability of a single-item measure based on the classical formula for correction for attenuation—we refer to the method as the CA method. Zijlmans et al. (2018) adjusted the MS method, Guttman's λ_6, and the LCRC method so that they can be used to estimate the reliability of a single-item measure. Below we have a brief summary for the MS and the LCRC methods. The formula for the extended Guttman's λ_6 in estimating reliability of item scores on a single item i (denoted as λ_{6i}) can be found in Appendix A. For details, please refer to the article by Zijlmans et al. (2018).

Both the MS and LCRC methods estimate item reliability by the ratio of true score variance and the total observed score variance under the framework of classical test theory, or, by

$$\rho_{ii'} = \frac{\sigma^2_{T_i}}{\sigma^2_{X_i}} = \frac{\sum_{x=1}^{m} \sum_{y=1}^{m} \left[\pi_{x(i),y(i')} - \pi_{x(i)}\pi_{y(i)}\right]}{\sigma^2_{X_i}}, \tag{1}$$

where i is the index for a single item, i' is an independent repetition of item i, x and y denote realizations of item scores (x, y $= 0, 1, \ldots, m$). X_i is the observed score on item i, and T_i is the true score on item i, which is the expectation of an individual's test score across independent repetitions. $\sigma^2_{T_i}$ is the true score variance on item i, and $\sigma^2_{X_i}$ is the observed score variance on item i; $\pi_{x(i),y(i')}$ is the joint cumulative probability of getting at least score x and at least score y on two independent repetitions, denoted by i and i'; $\pi_{x(i)}$ is the marginal cumulative probability of obtaining at least score x on item i; $\pi_{y(i)}$ is the marginal cumulative probability of obtaining at least score y on item i. The MS and LCRC methods use different approaches to estimate the joint cumulative probability $\pi_{x(i),y(i')}$. The MS method assumes a double monotonicity model and uses the mean of eight approximation methods (Molenaar & Sijtsma, 1988) as an estimate for the joint cumulative probability $\pi_{x(i),y(i')}$. The LCRC method assumes a latent class model and estimates the joint cumulative probability $\pi_{x(i),y(i')}$ by the probability to be in a particular latent class, and the probability of a particular item score given class membership. A key to obtain an accurate LCRC reliability estimate is to have a good idea of the number of latent classes. Compared to the CA method, the calculation for either the MS method or the LCRC method is more time consuming.

Zijlmans et al. (2018) conducted a simulation study to compare the performance of the CA method (Wanous & Reichers, 1996) with the modified MS, Guttman's λ_6, and LCRC methods in calculating single-item score reliability. In their standard condition, they simulated both dichotomous and polytomous scores for six items that are unidimensional. They found that the MS and CA methods were the most accurate in terms of bias and variability in estimating single-item reliability for each of the six items; the LCRC estimates had small bias, but large variation. In other conditions when they simulated 18 items that were not unidimensional, both the MS and CA methods produced larger bias; the MS estimates had larger bias than the CA estimates and the LCRC estimates; again, the LCRC estimates had small bias but large variation. Guttman's λ_6 always had the largest negative bias. The CA estimates seemed to be the best overall, and they are easy to compute.

According to existing literature, it is unclear which among the existing methods can be used to accurately estimate the reliability of the test whose design is presented in Table 1. The stratified alpha method, for instance, is not expected to lead to an accurate estimate of the reliability of scores on such tests because its formula requires the Cronbach's alpha for each section, and Cronbach's alpha cannot be computed for a section with a single item. While the Kristof and the Cronbach's alpha methods can be used to estimate the reliability of scores on tests with multiple sections where one of the sections includes only one item, the accuracy of the resulting reliability estimate is unknown based on past research. Specifically, the Cronbach's alpha is probably too low as an estimate for the internal consistency reliability of scores on tests with designs as presented in Table 1. Therefore, we need to explore or propose new methods to estimate reliability for such tests.

Based on the findings from van der Ark et al. (2011) and Zijlmans et al. (2018), we came up with the idea of combining the CA method with the stratified coefficient alpha method to estimate reliability of test scores on multisectional CR tests with a single-item section (as described in Table 1). We are interested in knowing how this two-step method performs compared to other feasible methods.

The Four Reliability Estimation Methods Compared in This Study

The two-step method we proposed is referred to as the CA-STR method and involves the use of the CA method (Wanous & Reichers,1996) to compute a reliability estimate for the section with only one item and then the use of the stratified coefficient alpha (Cronbach et al., 1965) to estimate reliability of test scores. One assumption of the CA method for estimating reliability of a single item is that the deattenuated correlation between the single item and the rest of the test is equal to 1. Method CA was chosen as the first component of our method because it performed better than Guttman's λ_6 and the LCRC method when the data are unidimensional based on Zijlmans et al. (2018), and it is easier to compute compared to the MS method.

The three other reliability estimation methods that we examined include the Cronbach's alpha (that is arguably the most popular reliability estimate in operational practice; see, for example, Raykov & Marcoulides, 2015), the Kristof reliability coefficient, and the stratified coefficient alpha after combining the single-item section with one of the other two sections of the test (Str_Combined). We proposed this Str_Combined method and included it in this study because of its simplicity (both in concept and calculation). The formulas and steps for computation for the four reliability estimates that we compared are given below.

The CA-STR Method

The CA-STR method: As described earlier, the first step in the CA-STR method is to estimate the reliability of the scores on the single-item section. Instead of using regular coefficient alpha in the denominator of Formula 15 (Zijlmans et al., 2018, p. 560), we used stratified coefficient alpha in the denominator. The reliability of the score on the single-item section was estimated as

$$\rho^{CA} = \frac{\rho_{UV}^2}{\rho_{VV'}^{Stratified}}, \tag{2}$$

where U is the score on the single-item section (e.g., Item 1 in Table 1), V is the score on all other items (e.g., Items 2–6 in Table 1), ρ_{UV}^2 is the squared correlation between scores on the single-item section and scores on all other items (e.g., Items 2–6 in Table 1), and $\rho_{VV'}^{Stratified}$ is the stratified coefficient alpha for scores on all other items (i.e., Items 2–6 in Table 1) that is computed as

$$\rho_{VV'}^{Stratified} = \left(1 - \frac{\sum \sigma_{V_j}^2 \left(1 - \alpha\rho_{V_jV_j'}\right)}{\sigma_V^2}\right), \tag{3}$$

where V is the total score on all sections except for the single-item section, $\sigma_{V_j}^2$ is the variance associated with total scores on section j that has more than one item, and $\alpha\rho_{V_jV_j'}$ is the Cronbach's alpha for section j that has more than one item.

Cronbach's Alpha

The formula for Cronbach's alpha $\alpha\rho_{VV'}$ is given by

$$\alpha\rho_{VV'} = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_{V_i}^2}{\sigma_V^2}\right), \tag{4}$$

where n is the number of items in a test, $\sigma_{V_i}^2$ is the variance associated with each item i, and σ_V^2 is the variance associated with total score.

Kristof Reliability Coefficient

This coefficient is computed using Formulas 9 and 13 in Kristof (1974) as

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} \tag{5}$$

and

$$\sigma_T^2 = \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} + \frac{\sigma_{12}\sigma_{23}}{\sigma_{13}} + \frac{\sigma_{13}\sigma_{23}}{\sigma_{12}} + 2\left(\sigma_{12} + \sigma_{13} + \sigma_{23}\right), \tag{6}$$

where X is the observed total score on all sections; σ_T^2 is the true score variance; σ_X^2 is the observed (total) score variance; ρ is the Kristof reliability coefficient; and σ_{mn}, $m \neq n$, $m = 1, 2, 3$; $n = 1, 2, 3$ is the covariance between the observed scores on section m and section n.

Table 2 Test Specifications for Operational and Pseudo Test 1 (30 Forms)

	Test 1			Pseudo Test 1		
Section	Number of items	Item sequence	Item score range	Number of items	Item sequence	Item score range
Section 1	5	1–5	0–2	5	1–5	0–2
Section 2	2	6–7	0–3	2	6–7	0–3
Section 3	3	8–10	0–4	1	8	0–4

Table 3 Test Specifications for Operational and Pseudo Test 2 (83 Forms)

	Test 2			Pseudo Test 2		
Section	Number of items	Item sequence	Item score range	Number of items	Item sequence	Item score range
Section 1	6	1–6	0–2	6	1–6	0–2
Section 2	6	7–12	0–2	6	7–12	0–2
Section 3	3	13–15	0–4	1	13	0–4

Str_Combined Method

In this method, the first step is to select a section to be combined with the single-item section. The second step is to compute stratified coefficient alpha for the test with two sections using Formula 3 above. Given the data we used in this study (Tables 2 and 3), we combined Section 3 with Section 2 before calculating stratified coefficient alpha for Section 1 and the combined section (denoted as Section 23).

Real Data Study

The real/operational data we used in this study were from two tests that will be referred to as Test 1 and Test 2 and included only CR items. We obtained data from 30 operational forms of Test 1 and created 30 pseudo forms for Test 1 by removing the last two items from each operational form. The sample sizes of these pseudo forms ranged from 161 to 2,316; the average sample size was 580. Similarly, we created 83 pseudo forms for Test 2 by removing the last two items from 83 operational forms of Test 2. The sample sizes of these pseudo forms ranged from 344 to 4,606, with an average of 1,195. In the following sections, Pseudo Test 1 (or Pseudo Test 2) refers to those pseudo forms created from Test 1 (or Test 2). All of our analyses were based on data from these pseudo forms.

Tables 2 and 3 present the specifications of the pseudo forms we created based on Tests 1 and 2, respectively. Each test includes three sections that vary in item types, task complexities, number of items, and score ranges. The third section of the actual tests includes three items. We removed the last two items to obtain the pseudo forms that we used in our analysis.

To examine whether the Kristof coefficient is sensitive to part-test length ratio, we randomly manipulated the proportion of each section score relative to the total test score for both tests by applying different weights to each section. Table 4 shows the weights that we applied to the section scores and the resulting part-test length ratios for both pseudo tests and both part-test ratio conditions. For pseudo test 1, the first part-test length ratio was 5:3:2 and the second part-test length ratio was approximately 1:3:5. For pseudo test 2, the first part-test length ratio was 3:3:1 and the second part-test length ratio was 1:1:1. For both pseudo tests, section 1 has a relatively larger contribution toward the total test score, while section 3 has a relatively smaller contribution toward the total test score in part-test length ratio 1 than in ratio 2.

Simulation Study

We conducted a simulation study to compare the CA-STR method with the other methods to examine the accuracy of the method in comparison to other reliability estimation methods. In the simulation study, we varied not only the part-test length ratios but also test dimensionality to examine how the relative performance of these four reliability estimation

Table 4 Part-Test Length Ratio (Proportion of Section Scores Relative to Total Score)

Test	Part-test	Max. section scores (see Tables 2 and 3)	Weights on section scores	Max. total score	Part-test length ratio
Pseudo Test 1	Part-Test Length Ratio 1	10, 6, 4	1, 1, 1	$10 + 6 + 4 = 20$	5, 3, 2
	Part-Test Length Ratio 2	10, 6, 4	1, 5, 12	$10 + 30 + 48 = 88$	1, 3, 5[a]
Pseudo Test 2	Part-Test Length Ratio 1	12, 12, 4	1, 1, 1	$12 + 12 + 4 = 28$	3, 3, 1
	Part-Test Length Ratio 2	12, 12, 4	1, 1, 3	$12 + 12 + 12 = 36$	1, 1, 1

[a] This is an approximation.

methods varies across different situations. The simulated data have the same test specifications as those of the aforementioned pseudo forms. The simulated data also have two different part-test length ratios, as in the case of the real data study. Both unidimensional and two-dimensional data (with correlation of 0.4 and 0.7 between two dimensions) were simulated.

The steps of the simulation study for either pseudo test 1 or pseudo test 2 are given by the following:

1. Calibrate each of the pseudo forms for the test used in the real data study using a unidimensional generalized partial credit model (GPCM; Muraki, 1992).
2. Pick one of the pseudo forms for the test with relatively good model fit and larger sample size. If several items in a data set had a statistically significant value of the generalized $S\text{-}X^2$ item fit statistic (Kang & Chen, 2008), then the model is assumed to not fit the data set well.
3. Use item parameters from the selected form and theta values generated from a standard normal distribution to generate item scores on the test for $n = 100{,}000$ test takers based on a GPCM. This data set is considered as the population data set for the selected form at Time 1.
4. Starting with the same sample of 100,000 theta values and the same item parameters, select a new random-number seed and simulate a new set of item scores for each of the thetas in the sample. Call this the population dataset at Time 2.
5. Calculate correlation between total test scores at Time 1 and Time 2. This correlation coefficient, which is a test-retest correlation, was considered as the true reliability coefficient for the selected form.
6. From the population data set at Time 1, draw 1,000 samples using the simple random sampling with replacement method. Three sample size conditions were considered: $n = 150$, $n = 300$, and $n = 1000$.[2]
7. For each sample drawn in Step 6, calculate total test reliability using the four methods—the resulting values constitute the sample reliability estimates.
8. Using the aforementioned true reliability and the sample reliability estimates, calculate the bias (difference between sample reliability estimates and true reliability coefficient) and the root mean squared error (RMSE) of the sample reliability estimates.

We then used a similar 8-step procedure with a revised first step to simulate two-dimensional data sets. In the revised Step 1, instead of fitting a unidimensional GPCM model, we fitted a two-dimensional GPCM (Reckase, 2009, p. 103). After evaluating model fit, item parameters from one pseudo form of Test 1 with acceptable model fit were used to simulate item scores based on a two-dimensional model. Most pseudo forms in Test 1 had bad model fit. None of the pseudo forms in Test 2 had acceptable model fit under two-dimensional GPCM; so, no multidimensional data simulation was conducted for Test 2. When generating item responses in Step 3, the theta values were simulated from a bivariate normal distribution with correlations of 0.4 or 0.7 between the two variables. The factor structure we used for generating the two-dimensional data is consistent with the factor structure in the selected pseudo form of Test 1. An exploratory factor analysis on the selected pseudo form shows that items from Section 1 correspond to one dimension and items from Sections 2 and 3 correspond to another dimension. Scores on Section 3 had higher correlation with scores on Section 2 than with scores on Section 1.

In this study, we used the GPCM (Muraki, 1992) to fit the real data and to generate item response scores because the GPCM is more often used in the field of educational testing than the graded response model (e.g., Bürkner et al., 2019). We used the open-source software R (e.g., R Core Team, 2022) for data simulation, calibration, model-fit assessment, and reliability computation.

As shown in Step 5, we used the test-retest reliability estimate based on the population data as the true reliability coefficient. We did not use any internal consistency reliability measure as the true reliability because we did not know of an appropriate internal-consistency measures for our particular tests yet. We computed the true reliability as the ratio of true score variance and observed score variance as in Raykov and Marcoulides (2015). The difference between the test-retest and the variance ratio reliability coefficients based on the population data is very small. For example, for the population data for Test 1, the differences between these two reliability coefficients were within 0.01 under the two part-test length ratios.

The RMSE in Step 8 is a summary statistic that reflects how accurate and stable each reliability estimate is. It is the square root of average squared difference between the estimated and the true value, across 1,000 samples. That is,

$$RMSE = \sqrt{\frac{\sum \left(\hat{\rho}_i - \rho\right)^2}{n}}, \tag{7}$$

where, n = 1,000, ρ is the true reliability coefficient, and $\hat{\rho}_i$ is the estimated reliability coefficient for sample *i*. The RMSE statistics takes both accuracy and variation of the estimation into account. An estimate with smaller RMSE is typically preferred over an estimate with larger RMSE.

van der Ark et al. (2011) interpreted absolute bias as follows: $|\text{bias}| < .001$ is considered negligible, $.001 \leq |\text{bias}| < .01$ is small, $.01 \leq |\text{bias}| < .02$ is medium, $.02 \leq |\text{bias}| < .05$ is considerable, and $|\text{bias}| \geq .05$ is considered large. In this study, we considered an absolute bias larger than .05 as large and an absolute bias $\leq .05$ as small. And we applied the same criterion to interpret the magnitude of RMSE.

Results

Results for Real Data

Tables 5 and 6 present a summary of various reliability estimates based on real data for Pseudo Tests 1 and 2 respectively. It should be noted that the true reliability is unknown for the real data; so we examined how the four reliability coefficients differed from each other. Table 5 indicates that under Part-Test Length Ratio 1 of Pseudo Test 1, the reliability estimates from the CA_STR and the Str_Combined methods were very similar, and they were the highest. The reliability estimates from the Kristof method were the lowest. However, when the proportion of scores in the three sections changed from Part-Test Length Ratio 1 to Part-Test Length Ratio 2 (6:3:2 to 1:3:5), the Kristof coefficients became the highest overall, and the regular coefficient alphas became the lowest. Table 6 indicates that for Pseudo Test 2, the CA_STR reliability estimates were the highest overall under both part-test length ratios. Reliability estimates from the Str_Combined method were

Table 5 Summary of Various Reliability Estimates Across 30 Pseudo Forms of Test 1 – Real Data

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
Statistic	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
Avg.	0.7699	0.6779	0.7934	0.7942	0.5822	0.7830	0.7771	0.6967
Min.	0.6153	0.5241	0.6266	0.6301	0.4187	0.6023	0.6031	0.4760
Max.	0.9021	0.8685	0.9187	0.9206	0.6877	0.9027	0.9027	0.8217

Note. Avg. = average; min. = minimum; max. = maximum.

Table 6 Summary of Various Reliability Estimates Across 83 Pseudo Forms of Test 2 – Real Data

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
Statistic	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
Avg.	0.7884	0.7960	0.8147	0.8007	0.7729	0.7942	0.8053	0.8010
Min.	0.7067	0.6950	0.7199	0.7184	0.6674	0.6946	0.7064	0.6915
Max.	0.8750	0.8920	0.9218	0.8955	0.8701	0.9008	0.9200	0.9159

Note. Avg. = average; min. = minimum; max. = maximum.

the second highest overall, and those from the Kristof method were the third highest. Reliability estimates based on the regular coefficient alpha method were the lowest under both part-test length ratios.

Results for Simulated Data

Table 7 shows the true reliability coefficients that were calculated as the correlations between scores from two simulated populational data sets.

Table 8 summarizes the different reliability estimates computed using the 1,000 simulated samples drawn from the unidimensional population at Time 1. Under all conditions, the Cronbach's alpha was always the lowest reliability estimate. For Pseudo Test 1, the highest reliability estimate was mostly the Kristof estimate. For Pseudo Test 2, the highest estimate was always the Kristof estimate. The CA_STR reliability estimate was very close to the Kristof estimate under all conditions. Table 8 also shows that all the reliability estimates were larger to some degree for Part-Test Length Ratio 1 compared to Part-Test Length Ratio 2. This result was consistent with the pattern shown by the true reliability coefficients in Table 7. Compared to the other estimates, the Cronbach's alpha coefficient had the largest decrease from Part-Test Length Ratio 1 to Part-Test Length Ratio 2. A visual representation of the summary of the reliability estimates can be found in Appendix B.

Table 7 True Reliability for Simulated Data

Test	Part-test ratio	Unidimensional	Two-dimensional with correlation 0.4	Two-dimensional with correlation 0.7
Test 1	Part-Test Ratio 1	0.6883	0.8380	0.8496
	Part-Test Ratio 2	0.6450	0.7808	0.7925
Test 2	Part-Test Ratio 1	0.8372	n/a	n/a
	Part-Test Ratio 2	0.8103	n/a	n/a

Table 8 Summary of Reliability Estimates Across 1,000 Simulated Samples — Unidimensional

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
				Pseudo Test 1				
n = 150	0.6685	0.6914	0.6821	0.6797	0.4924	0.6409	0.6410	0.5887
n = 300	0.6670	0.6863	0.6802	0.6784	0.4909	0.6378	0.6378	0.5874
n = 1,000	0.6687	0.6858	0.6816	0.6799	0.4911	0.6354	0.6355	0.5873
				Pseudo Test 2				
n = 150	0.8129	0.8309	0.8206	0.8141	0.7725	0.8045	0.7988	0.7845
n = 300	0.8131	0.8290	0.8205	0.8144	0.7720	0.8031	0.7984	0.7841
n = 1,000	0.8130	0.8277	0.8205	0.8143	0.7720	0.8023	0.7983	0.7842

Table 9 Average Reliability Estimation Bias Across 1,000 Simulated Samples — Unidimensional

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
				Pseudo Test 1				
n = 150	-0.0198	0.0031	-0.0062	-0.0086	-0.1526	-0.0041	-0.0040	-0.0563
n = 300	-0.0213	-0.0020	-0.0081	-0.0099	-0.1541	-0.0072	-0.0072	-0.0576
n = 1,000	-0.0196	-0.0025	-0.0067	-0.0084	-0.1539	-0.0096	-0.0095	-0.0577
				Pseudo Test 2				
n = 150	-0.0243	-0.0063	-0.0166	-0.0231	-0.0378	-0.0058	-0.0115	-0.0258
n = 300	-0.0241	-0.0082	-0.0167	-0.0228	-0.0383	-0.0072	-0.0119	-0.0262
n = 1,000	-0.0242	-0.0095	-0.0167	-0.0229	-0.0383	-0.0080	-0.0120	-0.0261

Tables 9 and 10 respectively provide the overall bias and RMSE of different reliability estimates across samples drawn from the unidimensional population. For Pseudo Test 1, under Part-Test Length Ratio 1, the regular coefficient alpha method was found to have underestimated the true reliability coefficient of Pseudo Test 1 by 0.0198 on average. Under Part-Test Length Ratio 2, the Cronbach's alpha method underestimated true reliability of Pseudo Test 1 even more, by 0.1526 on average. Table 9 shows that all the estimation biases were negative except for the Kristof coefficients when sample size is small (n = 150) under Part-Test Length Ratio 1 for Pseudo Test 1. This result was not surprising given that Sedere and Feldt (1977) also found that the Kristof coefficient can sometimes overestimate test-score reliability.

Table 10 shows the RMSE of each reliability estimation method under different simulation conditions. For Pseudo Test 1, the CA_STR method always had the smallest RMSEs. For Pseudo Test 2, the RMSEs of the CA_STR method were just slightly larger than those of the Kristof method, but quite smaller than the RMSEs of the other two methods.

The results from Tables 8–10 indicate that the Kristof coefficient was the highest reliability estimate under most conditions when data were unidimensional. Reliability estimates from the CA_STR method were very close to those Kristof estimates. When we examined the RMSEs that combined both estimation bias and estimation variation, the CA_STR reliability estimates seemed to be the best with smallest RMSEs under most conditions. Overall, the CA-STR method performed similar to or slightly better than the Kristof method when the data were unidimensional. In general, the regular coefficient alpha performed the worst among the four methods.

Tables 11–13 and 14–16 present results for data simulated from the two-dimensional GPCM with a correlation of 0.7 and 0.4, respectively, between the two latent ability variables. Item parameters used for this simulation were estimated from one form of Pseudo Test 1.

Table 10 Summary of RMSE Across 1,000 Simulated Samples—Unidimensional

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
				Pseudo Test 1				
$n = 150$	0.0473	0.0518	0.0432	0.0437	0.1573	0.0589	0.0581	0.0737
$n = 300$	0.0361	0.0361	0.0305	0.0308	0.1564	0.0421	0.0417	0.0668
$n = 1000$	0.0255	0.0209	0.0180	0.0185	0.1546	0.0245	0.0242	0.0607
				Pseudo Test 2				
$n = 150$	0.0321	0.0292	0.0275	0.0314	0.0443	0.0276	0.0277	0.0355
$n = 300$	0.0279	0.0215	0.0222	0.0269	0.0414	0.0197	0.0207	0.0309
$n = 1{,}000$	0.0255	0.0144	0.0187	0.0243	0.0392	0.0129	0.0152	0.0276

Table 11 Summary of Reliability Estimates Across 1,000 Simulated Samples—Two-Dimensional With Correlation 0.7

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
$n = 150$	0.8354	0.7795	0.8484	0.8478	0.5977	0.7677	0.7680	0.7020
$n = 300$	0.8358	0.7797	0.8487	0.8482	0.5985	0.7681	0.7686	0.7033
$n = 1{,}000$	0.8363	0.7790	0.8492	0.8488	0.5992	0.7681	0.7687	0.7043

Table 12 Summary of Reliability Estimation Bias Across 1,000 Simulated Samples—Two-Dimensional With Correlation 0.7

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
$n = 150$	-0.0142	-0.0701	-0.0012	-0.0018	-0.1948	-0.0248	-0.0245	-0.0905
$n = 300$	-0.0138	-0.0699	-0.0009	-0.0014	-0.1940	-0.0244	-0.0238	-0.0892
$n = 1{,}000$	-0.0133	-0.0706	-0.0004	-0.0008	-0.1933	-0.0244	-0.0237	-0.0882

Table 13 Summary of RMSE Across 1,000 Simulated Samples — Two-Dimensional With Correlation 0.7

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
n = 150	0.0264	0.0814	0.0213	0.0211	0.1965	0.0454	0.0440	0.0959
n = 300	0.0203	0.0752	0.0142	0.0141	0.1948	0.0362	0.0350	0.0918
n = 1,000	0.0158	0.0723	0.0082	0.0082	0.1935	0.0284	0.0276	0.0891

Table 14 Summary of Reliability Estimates Across 1,000 Simulated Samples — Two-Dimensional With Correlation 0.4

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
n = 150	0.8022	0.7039	0.8283	0.8275	0.5873	0.7707	0.7675	0.7032
n = 300	0.8033	0.7061	0.8292	0.8285	0.5886	0.7704	0.7684	0.7052
n = 1,000	0.8036	0.7045	0.8296	0.8289	0.5892	0.7695	0.7681	0.7060

Table 15 Summary of Reliability Estimation Bias Across 1,000 Simulated Samples — Two-Dimensional With Correlation 0.4

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
n = 150	-0.0358	-0.1341	-0.0097	-0.0105	-0.1935	-0.0101	-0.0133	-0.0776
n = 300	-0.0347	-0.1319	-0.0088	-0.0095	-0.1922	-0.0104	-0.0124	-0.0756
n = 1,000	-0.0344	-0.1335	-0.0084	-0.0091	-0.1916	-0.0113	-0.0127	-0.0748

The results in Tables 11–16 indicate that when the data were two-dimensional with either higher or lower correlations between the two dimensions, the performance of the Kristof method and the Cronbach's alpha method changed significantly when the part-test length ratio changed. The Kristoff reliability estimates were the smallest under Part-Test Length Ratio 1 with associated largest RMSEs but the largest under Part-Test Length Ratio 2 with associated smallest RMSEs (Tables 11, 13, 14, and 16). This result was in contrast with the result for unidimensional data that the RMSEs of the Kristof reliability estimates were nearly the smallest under all conditions for both Part-Test Length Ratios 1 and 2 (Table 10). The regular coefficient alpha reliability estimates were significantly larger for Part-Test Length Ratio 1 compared to Part-Test Length Ratio 2. The reliability estimates from the CA_STR method were almost always the highest and had the smallest RMSEs for two-dimensional data. In Table 11 (also refer to Figure B3 in Appendix B), all reliability estimates were larger for Part-Test Length Ratio 1 compared to Part-Test Length Ratio 2. In Table 14 (refer to Figure B4 in Appendix B), all reliability estimates were larger for Part-Test Length Ratio 1 than for Part-Test Length Ratio 2 except for the Kristof method. This pattern in Table 14 was also observed in the results based on real data from Pseudo Test 1 (Table 5).

Some of the findings in our study are consistent with those from previous studies. For example, first, in all three tables — Table 10, Table 13, and Table 16 — the RMSEs of all reliability estimates decreased as sample size increased. There was no such pattern for average bias. These results are consistent with the findings from van der Ark et al. (2011). Second, our results indicate that all the reliability estimates had negative bias under all conditions, except for the Kristof coefficient. The Kristof coefficients had negative bias in most cases, but a positive bias was observed in Table 9 when the simulated data is unidimensional with sample size of 150 under Part-Test Length Ratio 1 for Pseudo Test 1. Sedere and Feldt (1977) also found out that the Kristof coefficients can over-estimate test reliability. They found that sometimes, especially when sample size is small, the Kristof coefficients can be even greater than 1. Third, van der Ark et al. (2011) and Zijlmans et al. (2018) found that estimation bias was larger when data were two-dimensional than when data were unidimensional. The correlation between the two latent ability variables was 0 and 0.5 in the two studies. In our study, bias was larger under all conditions for two-dimensional data than for unidimensional data when the correlation between the two latent ability variables was 0.4. The same pattern held for Part-Test Length Ratio 2 when the correlation was 0.7. These consistencies support the reasonableness of our results.

Table 16 Summary of RMSE Across 1,000 Simulated Samples—Two-Dimensional With Correlation 0.4

	Part-Test Length Ratio 1				Part-Test Length Ratio 2			
N	Alpha	Kristof	CA_STR	Str_Combined	Alpha	Kristof	CA_STR	Str_Combined
$n = 150$	0.0445	0.1434	0.0255	0.0256	0.1952	0.0396	0.0379	0.0834
$n = 300$	0.0390	0.1365	0.0180	0.0184	0.1930	0.0296	0.0283	0.0787
$n = 1000$	0.0359	0.1349	0.0124	0.0129	0.1919	0.0189	0.0190	0.0758

Conclusions and Recommendations

In this paper, we aimed to find better ways to estimate the internal consistency reliability of scores on tests with a specific type of design that we encounter in practice—these are tests with CR items clustered into sections that are not parallel or tau-equivalent, and one of the sections has only one item that is of a different type from the other items on the test. We proposed a two-step approach (denoted as CA_STR) to estimate the reliability of scores on this kind of tests and compared the performance of the CA_STR method with three other reliability estimation approaches using both real and simulated data.

Our results indicate that the CA_STR method provides the most accurate and robust reliability estimates under almost all conditions. The method can be used for reliability estimation for tests similar to those considered in our study. The CA_STR method first uses the CA method to calculate single item reliability in the last section of our pseudo-tests and then uses the stratified coefficient alpha method to estimate reliability of the total test score. When estimating the reliability of the single item in the last section, we adjusted the formula for the CA method by using stratified coefficient alpha instead of regular coefficient alpha. Using .05 as the cut point between small and large absolute bias and RMSE, the CA_STR method was the only method that produced test reliability estimates with small absolute biases and RMSEs under almost all conditions. The only exception was that the RMSE of the CA_STR estimates was slightly larger than .05 (.0581 in Table 10) when sample size was 150 under Part-Test Length Ratio 2 for unidimensional data simulated based on Pseudo Test 1. The CA_STR estimates were least affected by the dimensionality of the test, or the relative importance of each section scores, or sample size. Based on our study, we can conclude that the CA_STR estimates can be adequately accurate when sample size is at least 150 for both unidimensional or two-dimensional CR tests with only one item in a section. The other reliability estimation methods (especially the Kristof method and the Cronbach's alpha) were more sensitive to dimensionality of the test or the relative importance of each section score.

In addition to its enhanced estimation accuracy and stability, the CA_STR method offers another benefit in that the method provides not only reliability estimates of scores on the whole test, but also reliability estimates for scores on a single item. Thus, the CA_STR method can be used to estimate reliability of section scores even when some sections have only one item. None of the other three methods can provide reliability estimates for scores in a section that includes only one item. This advantage can set the CA_STR method apart from the others, as researchers or test users sometimes require section-level reliability estimates for operational purposes.

Even though the single-item section was in the last section of the test in our study, the computation of the CA_STR method does not require a specific placement of the single-item section and the single-item section can be anywhere in a test.

The Str_Combined method performed better than Cronbach's alpha, but worse than the CA_STR method. Further research can be conducted to examine whether the Str_Combined method will perform better when the two combined sections are more or less content homogeneous compared to the data we had in this study. In this study, we combined Section 3 with Section 2 instead of Section 1 since Section 3 appeared to be more homogeneous with Section 2 than with Section 1 based on the results of factor analysis. We hypothesized that the Str_Combined method would perform better when the single-item section is combined with a more homogeneous section than with a less homogeneous section. In future studies, we plan to combine Section 3 with Section 1 and examine whether our hypothesis is correct.

As expected, the Cronbach's alpha coefficient is found to be not the best reliability estimate under any condition since our items were not parallel or essentially tau-equivalent. The Cronbach's alpha coefficient always has larger RMSE than the CA_STR method (Table 10, Table 13, Table 16).

Our results provide proof of the instability of the Kristof coefficients in estimating the reliability of scores on CR tests with only one item in a section. Our simulation study shows that in some conditions, the Kristof coefficients can have small estimation bias or RMSE, but in some other conditions, they can have large RMSEs. For example, when simulated data are unidimensional, and when sample size is at least 300, the Kristof coefficients had the smallest RMSEs under both part-test length ratios for Pseudo Test 2. But when data are two-dimensional, the Kristof coefficients had the largest RMSEs among all four estimation methods for one of the part-test length ratios. In Table 14, the Kristof coefficients were the smallest under Part-Test Length Ratio 1, but they increased significantly while all other estimates decreased when the part-test length ratio changed. The increasing pattern of the Kristof reliability estimates from Part-Test Length Ratio 1 to Part-Test Length Ratio 2 in Table 14 is consistent with the results based on real data for Pseudo Test 1 (Table 5). However, it is not consistent with the decreasing pattern shown by the true reliability coefficients in Table 7. This inconsistency with true test reliability implies the inaccuracy of the Kristof estimates when data are two-dimensional. Our study shows that the usefulness of the Kristof coefficient in estimating reliability of total scores is very limited. The Kristof coefficient cannot be used for estimating test score reliability if sample size is up to 300 or if the test is multidimensional. Sedere and Felt (1977) also found that the Kristof coefficient does not work well when sample size is small. In addition, this method cannot be used for estimating single-item reliability.

Under all conditions, the part-test length ratio substantially affected our results including the magnitude of the true reliability and the relative performance of the various reliability estimation methods. The true reliability was larger for both pseudo tests for the first part-test length ratio compared to the second (see Table 7). This result is related to how the reliability of a weighted composite score varies as the relative contribution of each section score to the total test score (part-test length ratios or weights in Table 4) changes. Kane and Case (2004) showed that placing larger weight on the more reliable scores tends to improve the reliability of the weighted composite score. The weights were indeed larger on the more reliable scores in the first part-test length ratio compared to the second for both pseudo tests.

Regarding the relative performance of the various reliability estimation methods, when we change the part-test length ratio, we also change the variance-covariance matrix among the section scores. A change in the variance-covariance matrix affects the performance of the reliability estimation methods in different ways. The Kristof coefficient and Cronbach's alpha seem to be more impacted by (or less robust to) the changes to the variance covariance matrix, while the CA_STR and Str_Combined methods are not. Both CA_STR and Str_Combined methods involve stratified coefficient alpha and Feldt and Qualls (1996) showed that stratified coefficient alpha is more accurate than regular coefficient alpha when the test is not content homogeneous. In addition, the first step of the CA_STR method involves a formula based on correlations, which is not affected by changing relative part-test length ratios. In contrast, the computation of Kristof coefficient and Cronbach's alpha is based on covariances among the parts—so these two coefficients are affected more by changing the part-test length ratio.

Though this paper suggests a potential solution to an important practical problem, it has several limitations and leaves considerable scope for further research. As Wanous and Reichers (1996) pointed out, it is important to remember that reliability estimates vary among samples and must be re-estimated with each new research study. In practice, the performance of these reliability estimation methods may vary across different test designs or factor structures (i.e., test blueprints, indicating the score categories of each item, number of items in each section, number of sections in each test, and item formats) and different samples (i.e., test-taking populations). More studies are needed to examine if the CA_STR reliability estimates still perform the best on other types of CR tests and samples. Will the CA_STR method still perform the best when more than one section in the test has only one item? Will the Str_Combined method perform better when the correlation between the two combined sections is higher? Future studies may also evaluate the feasibility of G-theory and the bias in G-theory-based estimates under situations similar to those in this study.

Notes

[1] Double monotonicity means monotonically increasing item response functions and nonintersecting response functions of different items.

[2] Population datasets at time 2 were not used to draw any random samples. They were only used to compute the true reliability coefficients.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Brennan, R. L. (2017). *Using G theory to examine confounded effects: "The problem of one"* (CASMA Research Report #51). Center for Advanced Studies in Measurement and Assessment.

Bürkner, P.-C., Schwabe, R., & Holling, H. (2019). Optimal designs for the generalized partial credit model. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 271 – 293. https://doi.org/10.1111/bmsp.12148

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297 – 334. https://doi.org/10.1007/BF02310555

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

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*(2), 291 – 312.

Feldt L. S. & Qualls A. L. (1996). Bias in coefficient alpha arising from heterogeneity of test content. *Applied Measurement in Education*, *9*(3), 277 – 286. https://doi.org/10.1207/s15324818ame0903_5

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255 – 283. https://doi.org/10.1007/BF02288892

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, *17*(3), 221 – 240. https://doi.org/10.1207/s15324818ame1703_1

Kang, T. & Chen, T. T. (2008). Performance of the generalized $S\text{-}X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391 – 406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, *39*(4), 491 – 499. https://doi.org/10.1007/BF02291670

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151 – 160. https://doi.org/10.1007/BF02288391

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, *9*(28), 115 – 126. https://www.vvsor.nl/wp-content/uploads/2020/06/KM1988028006.pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159 – 176. https://doi.org/10.1177/014662169201600206

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1 – 13. https://doi.org/10.1007/BF02289400

The R Development Core Team. (2022). R: A language and environment for statistical computing. *R Foundation*. https://ringo.ams.stonybrook.edu/images/2/2b/Refman.pdf

Raykov, T., & Marcoulides, G. A. (2015). A direct latent variable modeling based method for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement*, *75*(1), 146 – 156. https://doi.org/10.1177/0013164414526039

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Sedere, M. U., & Feldt, L. S. (1977). The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's Lambda-2. *Journal of Educational Measurement*, *14*(1), 53 – 62. https://doi.org/10.1111/j.1745-3984.1977.tb00029.x

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. https://doi.org/10.4135/9781412984676

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271 – 295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380 – 392. https://doi.org/10.1177/0146621610392911

Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports*, *78*(2), 631 – 634. https://doi.org/10.2466/pr0.1996.78.2.631

Zijlmans, E. A. O., van der Ark, L. A., Tijmstra, J. and Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied Psychological Measurement*, *42*(7), 553 – 570. https://doi.org/10.1177/0146621618758290

Appendix A

$$KR20 = \left(\frac{n}{n-1}\right)\left(\frac{\delta_X^2 - \sum_{i=1}^{n} p_i\left(1-p_i\right)}{\delta_X^2}\right),$$

Where X is the observed total score on the test,
δ_X^2 is the sample variance of the observed total score,
n is the number of items,
p_i is the proportion of test takers who answered item i correctly.
Guttman (1945) defined the following six λ values:

$$\lambda_1 = 1 - \frac{\sum_{i=1}^{n} \sigma_i^2}{\sigma_X^2},$$

where σ_i^2 is the sample variance of scores on item i.

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1} C_2}}{\sigma_X^2},$$

where C_2 is the sum of squares of the sample covariances between items for a given test.

$$\lambda_3 = \frac{n}{n-1}\lambda_1,$$

$$\lambda_4 = 2\left(1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_X^2}\right),$$

where σ_1^2 and σ_2^2 are the sample variances of scores on the first and second parts of a test.

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\overline{C}_2}}{\sigma_X^2},$$

let C_{2i} be the sum of the squares of the covariances of item i with other items; then $\overline{C}_2$ is the largest of C_{2i}.

$$\lambda_6 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sigma_X^2},$$

where e_i^2 is the variance of the errors of estimate of item i from its linear multiple regression on the remaining n - 1 items.

Zijlmans et al. (2018) extended λ_6 to estimate the reliability of a single item. We refer to the extended estimate as λ_{6i} that is computed as

$$\lambda_{6i} = \frac{\boldsymbol{\tau}_i'\left(\Sigma_{ii}\right)^{-1}\boldsymbol{\tau}_i}{\sigma_{X_i}^2},$$

where $\boldsymbol{\Sigma}_{ii}$ is the (n - 1) x (n - 1) inter-item variance-covariance matrix for the set of (n - 1) items that includes all items on the test except item i, $\boldsymbol{\tau}_i$ is a (n - 1) x 1 vector containing the covariance of item i with the other (n - 1) items, and $\sigma_{X_i}^2$ is the variance of the observed scores on item i.

Appendix B

Figures B1 to B4 summarize, using box plots, the reliability estimates from the condition with sample size of 300 in our simulation study.

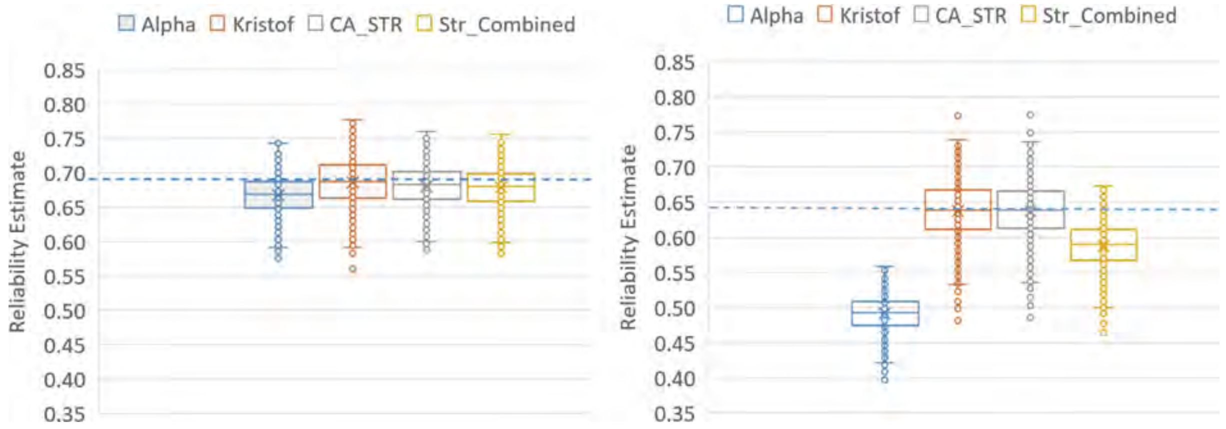

Figure B1 Reliability estimates and true reliability (dashed line). Unidimensional Test 1, Part-Test Length Ratio 1 (left) and Part-Test Length Ratio 2 (right). $n = 300$.

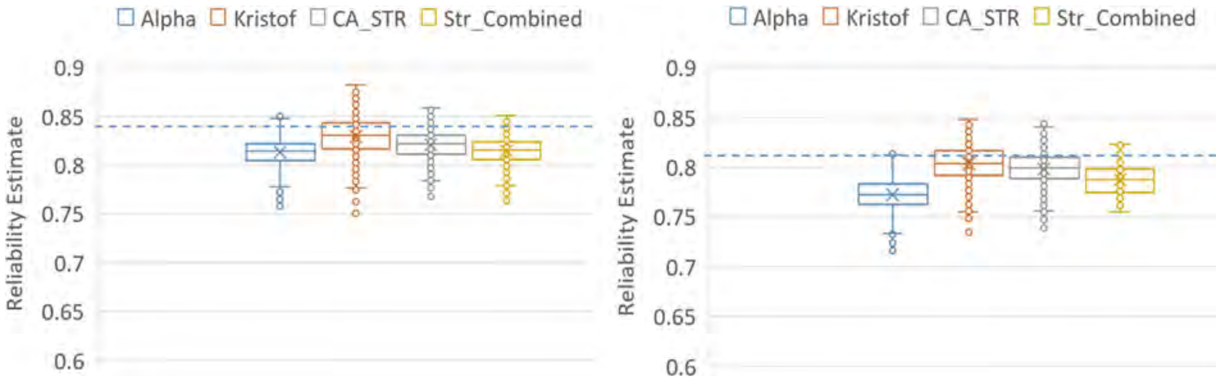

Figure B2 Reliability estimates and true reliability (dashed line). Unidimensional Test 2, Part-Test Length Ratio 1 (left) and Part-Test Length Ratio 2 (right). $n = 300$.

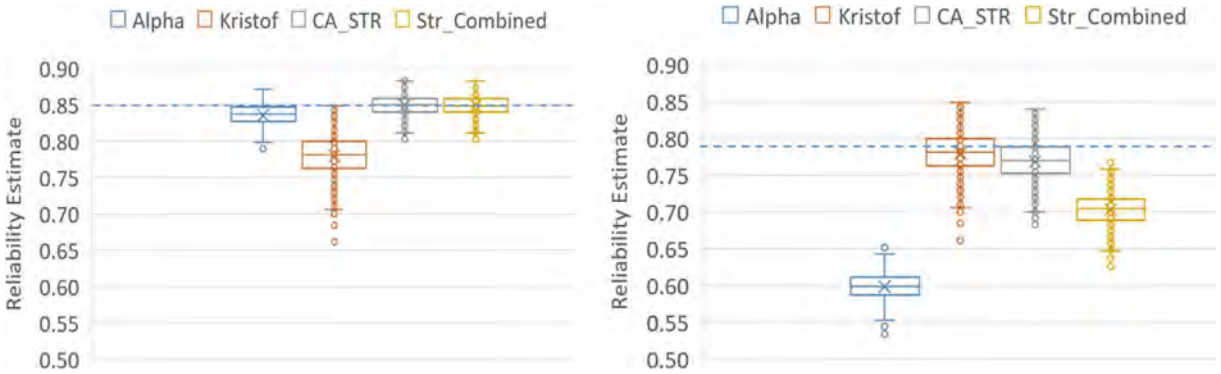

Figure B3 Reliability estimates and true reliability (dashed line). Multi-dimensional Test 1, Part-Test Length Ratio 1 (left) and Part-Test Length Ratio 2 (right). n = 300; correlation = 0.7.

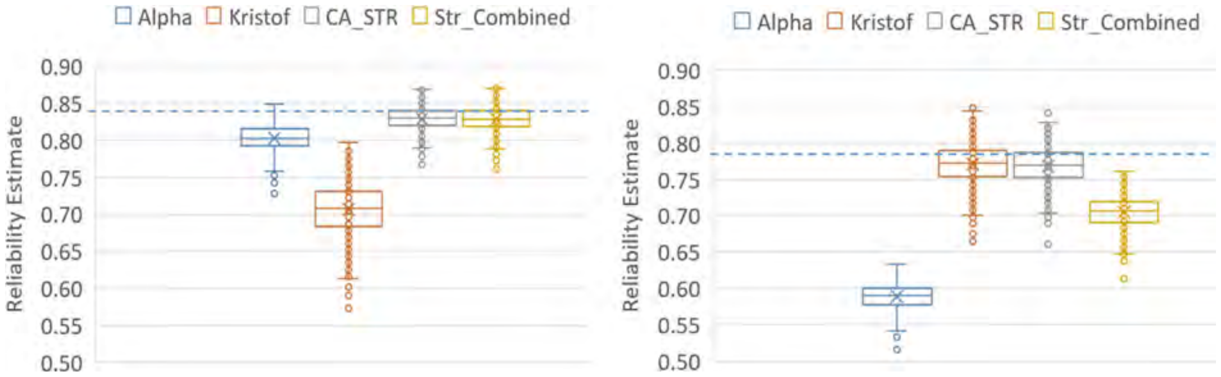

Figure B4 Reliability estimates and true reliability (dashed line). Multidimensional Test 1, Part-Test Length Ratio 1 (left) and Part-Test Length Ratio 2 (right). $n = 300$; correlation = 0.4.

Suggested citation:

Qu, Y., & Sinharay, S.. (2024). *Estimating reliability for tests with one constructed-response item in a section* (Research Report No. RR-24-07). ETS. https://doi.org/10.1002/ets2.12382

Action Editor: Daniel F. McCaffrey

Reviewers: Rui Gao and Jodi Casabianca-Marshall

Find other ETS-published reports by searching the ETS ReSEARCHER database.