Using Simulated Retests to Estimate the Reliability of Diagnostic Assessment Systems

Dr. W. Jake Thompson
Accessible Teaching, Learning, and Assessment Systems (ATLAS)
University of Kansas
1122 W. Campus Road, Lawrence, KS 66045
jakethompson@ku.edu
ORCiD: 0000-0001-7339-0300
Twitter: @atlas4learning

Brooke Nash
ATLAS
University of Kansas
1122 W. Campus Road, Lawrence, KS 66045
bnash@ku.edu
ORCiD: 0000-0001-9858-7062

Amy K. Clark
ATLAS
University of Kansas
1122 W. Campus Road, Lawrence, KS 66045
akclark@ku.edu
ORCiD: 0000-0002-5804-8336

Jeffrey C. Hoover
ATLAS
University of Kansas
1122 W. Campus Road, Lawrence, KS 66045
jhoover4@ku.edu
ORCiD: 0000-0002-0276-0308

Author Note

Correspondence concerning this manuscript should be addressed to W. Jake Thompson, ATLAS, University of Kansas, 1122 W. Campus Road, Lawrence, KS, 66045. Email: jakethompson@ku.edu.

Abstract

As diagnostic classification models become more widely used in large-scale operational assessments, we must give consideration to the methods for estimating and reporting reliability. Researchers must explore alternatives to traditional reliability methods that are consistent with the design, scoring, and reporting levels of diagnostic assessment systems. In this paper we describe and evaluate a method for simulating retests to summarize reliability evidence at multiple reporting levels. We evaluate how the performance of reliability estimates from simulated retests compares to other measures of classification consistency and accuracy for diagnostic assessments that have previously been described in the literature, but which limit the level at which reliability can be reported. Overall, the findings show that reliability estimates from simulated retests are an accurate measure of reliability and are consistent with other measures of reliability for diagnostic assessments. We then apply this method to real data from the Examination for the Certificate of Proficiency in English to demonstrate the method in practice and compare reliability estimates from observed data. Finally, we discuss implications for the field and possible next directions.

Keywords: diagnostic assessment, reliability, test-retest, simulation

Using Simulated Retests to Estimate the Reliability of Diagnostic Assessment Systems

Reliability of an assessment is a necessary and important source of validity evidence. Consistency of measurement must be demonstrated to support the valid interpretation and use of results. In the oft-given example, using a measuring tape to measure the length of a box should produce the same result each time. The same can be said of measurement in education. If a test is administered twice and provides accurate measurement of knowledge, skills, and understandings, the respondent should, in theory, receive the same score each time. This is the concept behind test-retest reliability (Guttman, 1945). Instances in which scores vary from one administration to the next indicate that the assessment lacks precision and that results are conflated with measurement error, which has an obvious negative impact on the validity of inferences made from the results.

In large-scale standardized testing environments, it is often impractical to administer the same assessment twice. Retest estimates may also be attenuated if knowledge is not retained between administrations or inflated if a practice effect is observed. For these reasons, reliability methods for operational programs often approximate test-retest reliability through other means. For example, Cronbach's (1951) coefficient alpha is one of the most commonly reported metrics of reliability for educational assessments. Rather than administering a test over two occasions, as is done for test-retest reliability, coefficient alpha determines the average of all the possible split-half reliability calculations for the assessment and represents the ratio of true score variance to observed score variance, effectively treating the halves as separate forms administered at the same time.

Selection of a method for estimating the reliability of an assessment depends on several factors, including the design of the assessment, the scoring model used to provide results, and the availability of data. The guidelines put forth by the *Standards for Educational and Psychological Testing* (*Standards* hereafter; American Educational Research Association [AERA] et al., 2014) specify a number of considerations for reporting reliability of assessment results. For the purposes of this paper, we focus on three specific standards:

- Standard 2.2: “The evidence provided for the reliability/precision of the scores should be consistent with the domain of replications associated with the testing procedures, and with the intended interpretations for use of the test scores” (p. 42).
- Standard 2.3: “For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported” (p. 43).
- Standard 2.5: “Reliability estimation procedures should be consistent with the structure of the test” (p. 43).

Because classical test theory (CTT) and item response theory (IRT) models have dominated the field of educational measurement, methods for evaluating reliability aligned to these models have similarly dominated the reliability literature (Brennan, 2001; Haertel, 2006). While methods of obtaining traditional reliability estimates are well understood and documented, there is far less research on methods for calculating the reliability of assessment results derived from less commonly applied statistical models, namely, diagnostic classification models (DCMs).

Diagnostic Classification Models

DCMs, also known as cognitive diagnosis models (CDMs; e.g., Leighton & Gierl, 2007), are confirmatory latent class models that represent the relationship of observed item responses to a set of categorical latent variables (e.g., Bradshaw, 2016; Rupp et al., 2010). Whereas traditional psychometric models (e.g., IRT) model a single, or occasionally multiple, continuous latent variables, DCMs model respondent mastery on a number of discrete latent variables (i.e., skills). Thus, a benefit of using DCMs for calibrating and scoring operational assessments is their ability to support instruction by providing fine-grained reporting at the individual skill level.

To provide detailed profiles of respondent mastery of skills measured by the assessment, DCMs require the specification of an item-by-skill (also referred to as item-by-attribute) matrix known as the Q-matrix (Tatsuoka, 1983). Based on the collected item-response data, the model determines the overall probability of respondents being classified into each latent class. The latent classes for DCMs are typically binary mastery status (master or nonmaster). This base-rate probability of mastery (i.e., the structural parameter) is then related to respondents' individual response data to determine the respondents' posterior probability of mastery for each assessed skill. The posterior probability is on a scale of 0 to 1 and represents the certainty the respondent has mastered each skill. Values closer to the scale extremes of 0 or 1 indicate greater certainty in the classification; a value of 0 indicates the respondent has definitely not mastered the skill, and a value of 1 indicates the respondent definitely has mastered the skill. In contrast, values closer to .50 represent maximum uncertainty in the classification. A mastery probability of .50 indicates the model cannot distinguish whether, on the basis of the available response data, the respondent has mastered the skill; the respondent is just as likely a master as a nonmaster. Diagnostic assessment results are typically reported as the mastery probability values or as dichotomous mastery statuses when a threshold for demonstrating mastery is imposed (e.g., .80). The dichotomous mastery statuses can also be aggregated into an skill mastery profile for reporting results.

The diagnostic scoring approach is unique in that the probability of mastery provides an indication of error or, conversely, certainty, for each skill and examinee. However, it does not provide information about consistency of measurement for the skill or for the assessment as a whole. Furthermore, because assessment results are the collection of skill-mastery results, rather than a total raw or scale score, traditional approaches to reliability are not appropriate, and alternate methods must be considered for reporting the reliability of operational assessment results.

Measuring the Reliability of Diagnostic Assessments

Because DCMs have not been widely used in operational or applied settings (Ravand & Baghaei, 2020; Sessoms & Henson, 2018), there has been limited research examining how best to report the reliability of classifications from a DCM-based assessment. However, there has been recent theoretical research on reliability methods for DCMs (for a review, see Sinharay & Johnson, 2019). In general, this research has been divided into two segments, depending on how results for the assessment are intended to be reported. If results are reported as the probability of mastery for each skill, then reliability should be reported as the precision of the estimated probability (e.g., Johnson & Sinharay, 2020; Templin & Bradshaw, 2013). In contrast, when results are reported as a binary classification (i.e., master or nonmaster) at the skill level, reliability is conceptualized as classification consistency and classification accuracy. This classification-based reliability will be the focus of this paper.

Classification accuracy is defined as the probability that an examinee receives a classification that is consistent with his or her true mastery status. *Classification consistency* is defined as the probability that an examinee receives the same classification across multiple administrations of an assessment (Cui et al., 2012). While classification-based reliability in

DCMs can be evaluated at multiple levels (e.g., skill level, profile level; Cui et al., 2012; Johnson & Sinharay, 2018; Wang et al., 2015), the definitions for classification accuracy and consistency are not altered by the level of analysis. Thus, the same statistical procedures can be used to estimate reliability in DCMs at each of these levels.

Early research on reliability in DCMs was conducted by Cui et al. (2012). They defined the cognitive diagnostic classification accuracy index and the cognitive diagnostic classification consistency index classification accuracy at the profile level. These indices provide the marginal probability of classifying an examinee accurately and consistently, respectively, at the profile level (Cui et al., 2012). However, these indices do not allow for evaluating accuracy and consistency at the skill level.

For assessments reporting results at the skill level, reliability evidence at the skill level should also be reported (e.g., Standard 2.3; Standard 2.5; Sinharay & Haberman, 2009). Wang et al. (2015) extended the work of Cui et al. (2012) by defining classification accuracy and consistency indices at the skill level. Wang et al. (2015) calculated skill-level classification accuracy and consistency as the proportion of examinees classified accurately and consistently within each skill's mastery status (i.e., masters and nonmasters).

While the classification accuracy and consistency indices defined by Wang et al. (2015) allow for calculating classification-based reliability at the skill level, Johnson and Sinharay (2018) noted these indices rely on the assumption that the posterior probabilities are constant across parallel forms of a test. Using a simple counterexample, Johnson and Sinharay demonstrated that this assumption is easily violated. They defined modified skill-level classification accuracy and consistency indices at the skill level using consistent estimators, and they provided interpretive guidelines for these new indices. Other commonly reported indices that Johnson and Sinharay suggested calculating include Youden's (1950) statistic, Goodman and Kruskal's (1954) lambda, Cohen's (1960) kappa, the tetrachoric correlation (Pearson, 1900), and sensitivity and specificity (Yerushalmy, 1947) to estimate reliability in DCMs.

Limitations of Current Classification-Based Reliability

The classification-based reliability indices defined by Cui et al. (2012), Wang et al. (2015), and Johnson and Sinharay (2018) can be calculated using data from a single administration, which acknowledges limitations pertaining to administering large-scale assessments multiple times. However, the existing classification-based reliability indices are limited to reporting reliability evidence at the skill and profile levels. This limitation may be problematic if results are aggregated and reported at a different level. For example, results may be reported as the total number of skills mastered or aggregated into an overall performance level (e.g., for state accountability systems) or pass/fail determinations (e.g., certification and licensure), yet the existing classification-based reliability indices do not support reporting reliability evidence at these levels. Thus, there is a need for methods to calculate classification-based reliability that are flexible for reporting multiple levels of reporting to support evidence recommended by Standards 2.2, 2.3, and 2.5 (AERA et al., 2014).

Simulation-Retest Reliability

Roussos et al. (2007) explained how simulated data obtained from calibrated DCM parameters (according to real data) can be used to produce summary statistics for evaluating a model, including several types of reliability indices. Specifically, the proportion of times each examinee is classified correctly for each skill was also described as providing an estimate of the correspondence between the estimated skill classification in the observed and simulated data. Similarly, the proportion of times each examinee is classified to the same category (e.g., masters or nonmasters) across two parallel tests was described as providing an estimate of test-retest consistency.

Templin and Bradshaw (2013) conducted a research study using a hypothetical second test administration to compare reliability estimates from a DCM to those of an IRT model for the same set of data collected from a single, fixed-form assessment administered to approximately 2,300 students. Rather than using a diagnostic assessment constructed with the purpose of reporting results at the skill level, this application retrofitted a DCM to existing large-scale assessment data designed to measure a single construct so that the assignment of items to skills was imposed post hoc. The researchers used posterior probabilities of mastery to calculate the probability of being assigned to each mastery profile and compared these probabilities to random draws from the theta distribution for the IRT-scored assessment. Reliability results comparing the mastery statuses obtained from the DCM were reported with a tetrachoric correlation for each skill in the model. While their main findings demonstrated that the DCM produced higher reliability estimates than those obtained from the IRT model for a test of the same length, they also demonstrated that hypothetical retest methods may be useful for evaluating reliability.

To report reliability evidence at multiple levels, a simulation-retest methodology is one method for evaluating reliability of diagnostic assessment results. Conceptually, a second administration of an assessment can be simulated on the basis of the administered assessment. By simulating a second administration, scores from two assessments are available, providing a means for evaluating retest reliability in the traditional sense (i.e., consistency of scores across multiple administrations). The simulation-retest approach differs from other CTT methods that report an estimate of the correlation between total scores from two forms, administrations, or halves of a test. Instead, a simulation-retest approach reports the correspondence between the estimated mastery statuses in the observed and simulated data, and the interpretation of the reliability results remains the same as for CTT methods. That is, reliability estimates are provided on a metric of 0 to 1, with values of 0 being perfectly unreliable and all variation attributed to measurement error, and values of 1 being perfectly reliable and all variation attributed to respondent differences on the construct measured by the assessment.

Consistent with existing classification-based reliability procedures, the simulation-retest methodology can be used to estimate the classification accuracy and consistency between the observed and simulated data at the skill and profile levels. However, the simulation-retest methodology also allows for estimating reliability for other aggregated reporting levels. Using the number of skills mastered to illustrate, it is possible to compare, for example, overall performance level in the observed and simulated administrations, and the reliability indices can be calculated to compare the consistency of the performance level determination. Similarly, the simulation-retest methodology can be used to estimate reliability at other levels of reporting.

Thompson et al. (2019) demonstrated how a simulation-retest method could be used to estimate the reliability of assessments scaled with DCMs at different levels of reporting. Thompson et al. applied the simulation-retest method to provide reliability evidence at multiple levels of reporting that are used for an operational, large-scale state assessment. The purpose of the current paper is to provide an in-depth description of the simulation-retest method for estimating reliability and compare the results from applying the simulation-retest method to those from other existing nonsimulation-based methods. Because existing nonsimulation-based methods cannot report reliability evidence at the all levels that results may be reported when mastery results are aggregated, the simulation-retest method offers a means for reporting reliability evidence and results at the same level. Consequently, it is important to compare the reliability estimates from the simulation-retest and nonsimulation-based methods at the skill level to generally demonstrate the accuracy and consistency of the simulation-retest method.

Calculating Reliability Estimates

The general approach to the simulation-retest reliability method, as described by Thompson et al. (2019), is to simulate a second set of responses based on actual respondent performance and calibrated-model parameters, score real-test data and simulated-test data, and compare respondents' estimated mastery statuses for the observed and simulated data. That is, once response data has been collected, calibrated, and scored, a second administration can be simulated using the known model parameters from the first (i.e., real) administration. In the context of using DCMs to calibrate and score the assessment, respondent performance is the set of mastery statuses for each skill. The threshold for mastery status must be specified before calculating reliability.

When calculating skill-level classifications, a threshold is specified to distinguish masters and nonmasters, recognizing that values farther from .50 indicate greater certainty in the classification. In applications of this methodology, the threshold value may vary depending on the design of the assessment, respondent population, stakeholder feedback, or other factors.

Applying the mastery threshold to the posterior probabilities of mastery obtained from the diagnostic scoring model results in a dichotomous mastery status for each skill measured by the assessment. The mastery status is one level of reporting results for diagnostic assessments and, therefore, one level at which reliability should be summarized. Because the scoring model produces mastery decisions, the term "results" is used instead of the term "scores" throughout this paper.

The specific steps for a DCM-based simulation are as follows:

1. Sample respondent record: Sample with replacement a respondent record from the operational data set. The respondent's mastery status or posterior probability of mastery from the operational scoring for each measured skill serves as the true value for the simulated respondent.
2. Simulate second administration: For each item the respondent was administered, simulate a new response that is based on the model-calibrated parameters, conditional on the true mastery probability or status for the skill.
3. Score simulated responses: Using the operational scoring method, assign mastery status by imposing a threshold for mastery on the posterior probability of mastery obtained from the model.
4. Repeat: Repeat the steps for a predetermined number of simulated respondents.

Step 1 draws respondent records from the operational data, and Step 2 simulates a second administration. This process ensures the simulation-retest method replicates results from real examinees using the actual set of items each examinee has taken, which means that the two administrations are perfectly parallel. In Step 3, the operational scoring procedure is applied to both the observed and simulated response data to calculate the posterior probability of mastery.

To calculate reliability indices, the estimated skill-mastery statuses for the observed and simulated data are compared across all replications determined in Step 4. Specifically, for each skill, reliability results are calculated using the 2×2 contingency table of estimated mastery statuses from the observed and simulated data, as shown in Table 1. We focus on skill-level reliability estimates in this paper because the nonsimulation-based methods are limited to reporting reliability evidence at the skill and profile levels; however, the benefit of the simulation-retest method is that the same procedure can be used for other levels of reporting. For example, we could calculate a performance level for both the observed and simulated data using an assessment’s operational rules. If there were four performance levels, we would then create a 4×4 contingency table similar to Table 1, showing the observed and simulated performance levels.

Table 1

2×2 Contingency Table of Estimated Mastery in the Observed and Simulated Administrations

Observed mastery status	Simulated mastery status	
	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

Note. 0 = skill nonmastery; 1 = skill mastery.

In this study, the performance of the simulation-retest reliability method is evaluated by comparing reliability estimates from the simulation-retest method with multiple nonsimulation-based reliability indices across a variety of simulated conditions. In addition to evaluating the simulation-retest reliability method through a simulation study, we also applied the simulation-retest method in an empirical data analysis of the grammar subtest of the Examination for the Certificate of Proficiency in English (ECPE; Templin & Hoffman, 2013), which was previously used by Sinharay and Johnson (2019) to demonstrate the application of a variety of classification-based reliability indices for DCMs.

Simulation Study

We conducted a simulation study to evaluate the accuracy of the reliability estimates from the simulation-retest method described above. In this study, we manipulated the number of assessed skills (three, four, five), the minimum number of items measuring each skill (three, four, five), the base rate of mastery (.10, .50, .90), the correlation between the assessed skills

(0.0, .35, .70), and item discrimination (low, moderate, high). This simulation used a full factorial design, resulting in 243 total conditions with 100 repetitions per condition.

Data Simulation

The simulation study is modeled on Johnson and Sinharay's (2018) evaluation of skill-level classification reliability indices. In the simulation for this study, each simulated assessment measured three, four, or five skills. The number of items included in each assessment (I) is the product of the number of assessed skills (A) and the minimum number of items measuring each skill (J; i.e., $I = A * J$). The Q-matrix (Tatsuoka, 1983) is specified so that the first six items form an identity matrix, and each remaining item has a 50% chance of assessing a second skill in addition to the identity matrix. Consistent with Johnson and Sinharay (2018), the items could not measure more than two skills. Table 2 presents an example Q-matrix for an assessment measuring three skills with a minimum of three items per skill.

Table 2

Example Q-Matrix

Item	Skill 1	Skill 2	Skill 3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	1	0	0
8	0	1	1
9	0	1	1

The base rate of mastery and the distributions for the item parameters were also simulated according to the approach used by Johnson and Sinharay (2018). The base rate of mastery for the first assessed skill was determined by the simulation condition, where 10%, 50%, or 90% of examinees mastered the first skill. The base rates of mastery for the remaining skills were determined by drawing a random number from a uniform distribution ranging from 0.2 to 0.8.

The generating model for this simulation was a log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), meaning the item parameters include item intercepts, main effects, and interaction effects. The item intercepts, which correspond to the probability of a nonmaster correctly responding to the item, were drawn from a uniform distribution ranging from 0.00 to 0.35, following the approach used by Johnson and Sinharay (2018). The item main effects, which correspond to the log odds increase in the probability for a master of the skill correctly responding to the item, were drawn from a truncated normal distribution with a mean of 1.0, 1.5, or 2.0 (representing low, moderate, and high discrimination, respectively) and a standard deviation of .17, where the values were constrained to be positive, using Johnson and Sinharay's (2018) approach. The item interaction effects, which correspond to the log odds increase in the probability for a master of two skills correctly responding to the item, were also drawn from a truncated normal distribution with a mean of 1.0, 1.5, or 2.0 and a standard deviation of .17, but the values were constrained to be greater than negative one times the smallest item main effect (i.e., -1 × min[main effects]) to meet the monotonicity constraints of the LCDM (Henson et al., 2009). Like the other item parameters, the distribution for the item-interaction-effect parameters followed the approach used by Johnson and Sinharay (2018).

In this study, 2,000 respondents were simulated for each generated data set. For each generated data set, we fit an LCDM and a deterministic-input, noisy-and-gate (DINA; de la Torre & Douglas, 2004; Junker & Sijtsma, 2001) model to each of the simulated data sets. Because the generating model for each data set is an LCDM, it is expected that the LCDM should demonstrate better fit than the DINA model. It is expected that these differences in model fit should have implications for classification-based reliability. Specifically, the DINA model is expected to demonstrate lower reliability because model misfit is present.

The generated data sets and the estimated model parameters were then used to create the simulated retests. When calculating the simulation-retest reliability estimates, 100,000 respondents were drawn with replacement and simulated for the retest data. A threshold of .50 was used in this simulation study to determine skill mastery, as this is a commonly used threshold in the literature (e.g., Bradshaw & Levy, 2019; Templin & Bradshaw, 2013). However, any threshold can be used in an operational setting. The simulation-retest classification consistency was then calculated as the proportion of respondent mastery classifications for each skill that matched the respondent's mastery status estimated from the original generated data set (i.e., how consistent the classifications were across the resampled respondents' simulated retests). Similarly, the simulation-retest classification accuracy for each skill was calculated as the average probability associated with each mastery classification across all simulated retests for the resampled respondents.

Method Comparisons

The simulation-retest reliability estimates for the LCDM were then compared to nonsimulation-based methods for estimating the reliability of DCMs. Specifically, the simulation-retest classification consistency was compared to the $\hat{P}_{ck}$ classification consistency measure defined by Johnson and Sinharay (2018; their equation 27). The simulation-retest classification accuracy was compared to the $\hat{\tau}_k$ measure defined by Wang et al. (2015; their equation 11). This measure was denoted as $\hat{P}_{ak}$ by Johnson and Sinharay (2018; their equation 9).

Because the simulated retests use the estimated model parameters to simulate item responses for the resampled respondents, it is implied that the estimated parameters are correct. Thus, it is possible that the simulation-retest method may produce biased estimates of reliability if there is model misfit. Therefore, it is important to examine the impact of model misfit on reliability estimates derived from the simulation-retest method. To evaluate the impact of model misfit, we simulated retests using the true data-generating parameters, the parameters estimated by the LCDM, and the parameters estimated by the DINA model. The LCDM was the data-generating model; therefore, the estimated LCDM parameters should be similar to the true parameters with some sampling variability. In contrast, the DINA model is a more restrictive model and therefore represents a model that does not truly fit the data and may therefore potentially bias the reliability estimates. Estimating both the LCDM and the DINA models allowed us to evaluate how reliability measures derived from simulated retests with parameters that either fit (i.e., the LCDM parameters) or did not fit (i.e., the DINA parameters) compared with the reliability measures derived from the true data-generating parameters. For all comparisons, we used the mean absolute difference to evaluate discrepancies between the reliability measures.

Simulation Study Results

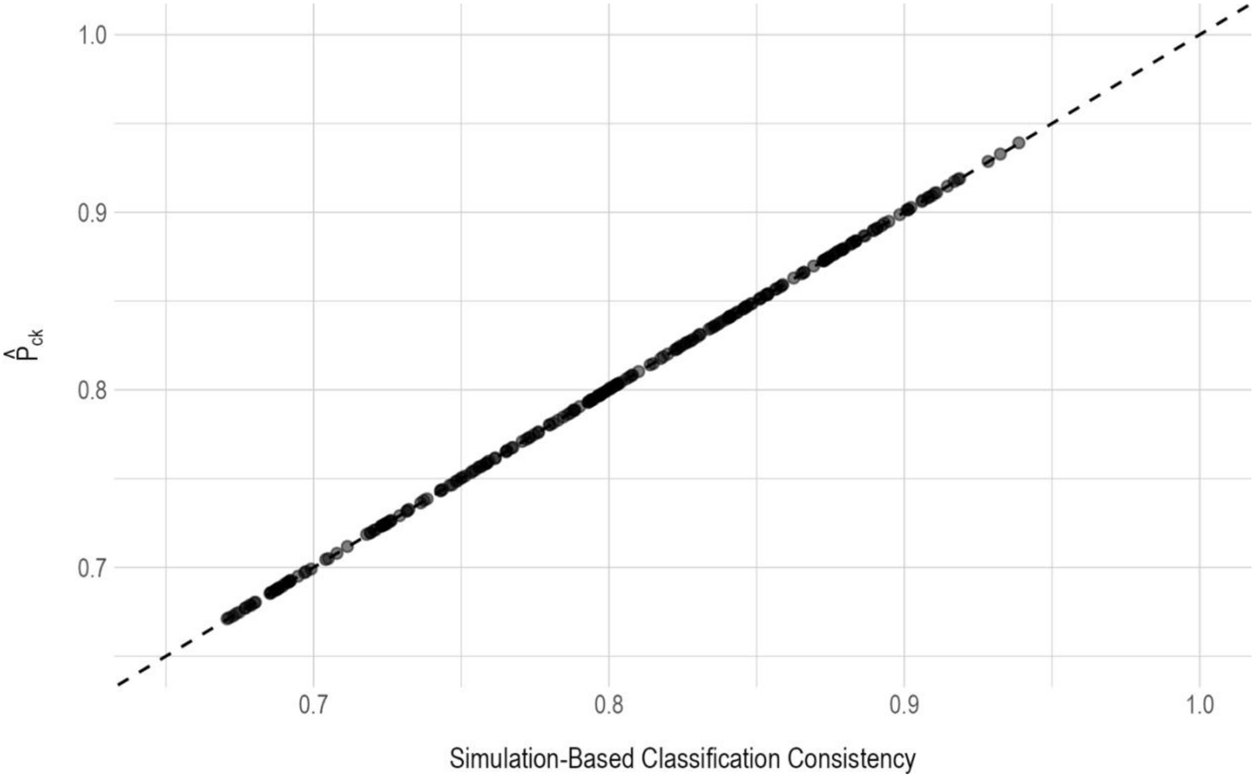
There were 243 conditions with 100 repetitions per condition in this simulation. The models and reliability estimates were estimated for 91% of repetitions using the true data-generating parameters as well as the estimated parameters from the LCDM and DINA models. The 9% of replications in which reliability estimates could not be estimated were evenly distributed across conditions; they were not the result of a limitation of the reliability method, but rather of the failure to converge of one or more of the estimated models. For simplicity, we report the estimated reliability for Skill 1 rather than for all skills that were included in each condition.

Classification Consistency

Figure 1 shows the average simulation-retest classification consistency and nonsimulation-based classification consistency ($\boldsymbol{\widehat{P}_{ck}}$; Johnson & Sinharay, 2018) for the first skill in each condition. Overall, the estimates from the simulation-retest and nonsimulation-based methods are highly consistent, with an average absolute difference of only 0.0002. The similarity between the two measures of classification consistency was stable across all simulation conditions, indicating that the simulation-retest measure of classification consistency provides reliability estimates comparable to nonsimulation-based measures across a variety of assessment conditions.

Figure 1

Comparison of Classification Consistency Across All Simulation Conditions



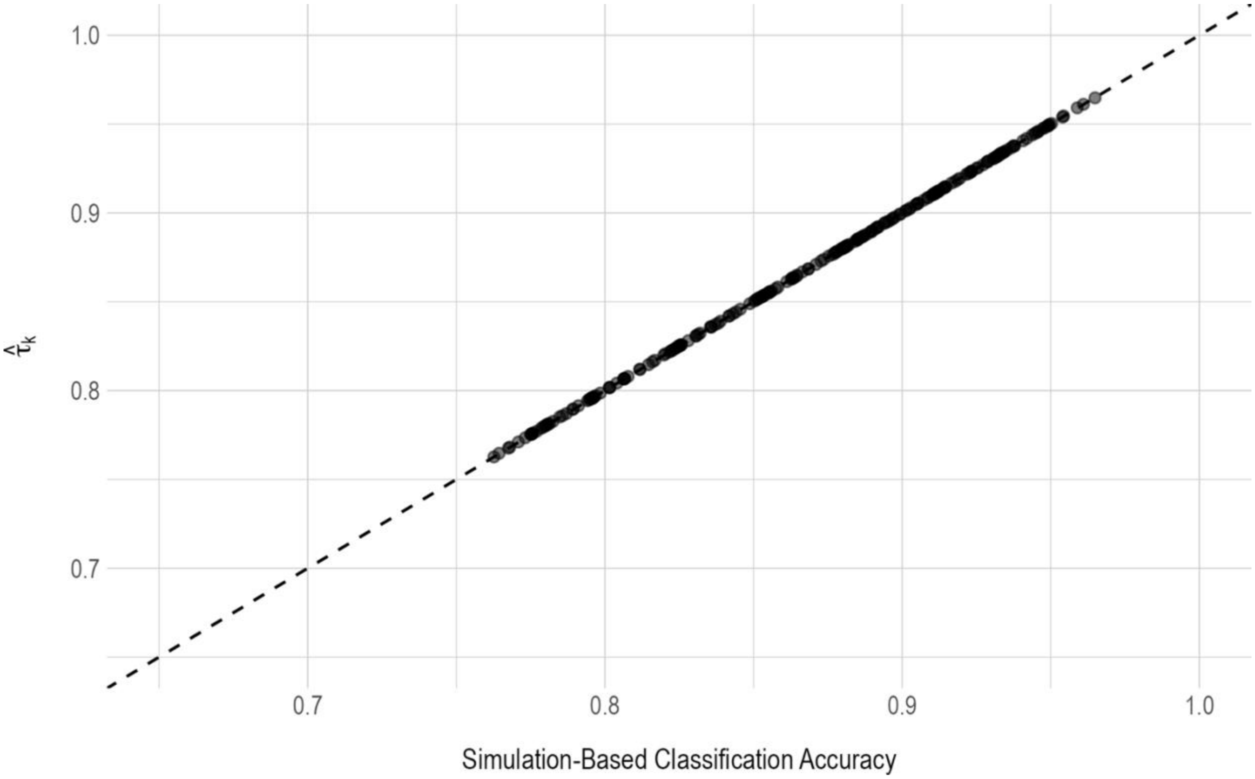
Note. Dashed line represents perfect agreement.

Classification Accuracy

When comparing the simulation-retest reliability estimates from the LCDM model with the nonsimulation-based reliability accuracy estimates, the classification accuracy estimates were also highly similar. Figure 2 shows a scatterplot with the average simulation-retest classification accuracy estimate for the first skill for each condition on the *x*-axis and the average nonsimulation-based classification accuracy estimate ($\hat{\tau}_k$; Wang et al., 2015) for each condition on the *y*-axis. In the scatterplot, the dashed line is the line of perfect agreement. The simulation-retest and nonsimulation-based reliability estimates are close to the line of perfect agreement, with an average absolute difference of 0.0001 across conditions. Thus, as with the classification consistency, when the estimated model matches the generating model, simulation and nonsimulation-based methods give nearly identical estimates of skill reliability.

Figure 2

Comparison of Classification Accuracy Across All Simulation Conditions



Note. Dashed line represents perfect agreement.

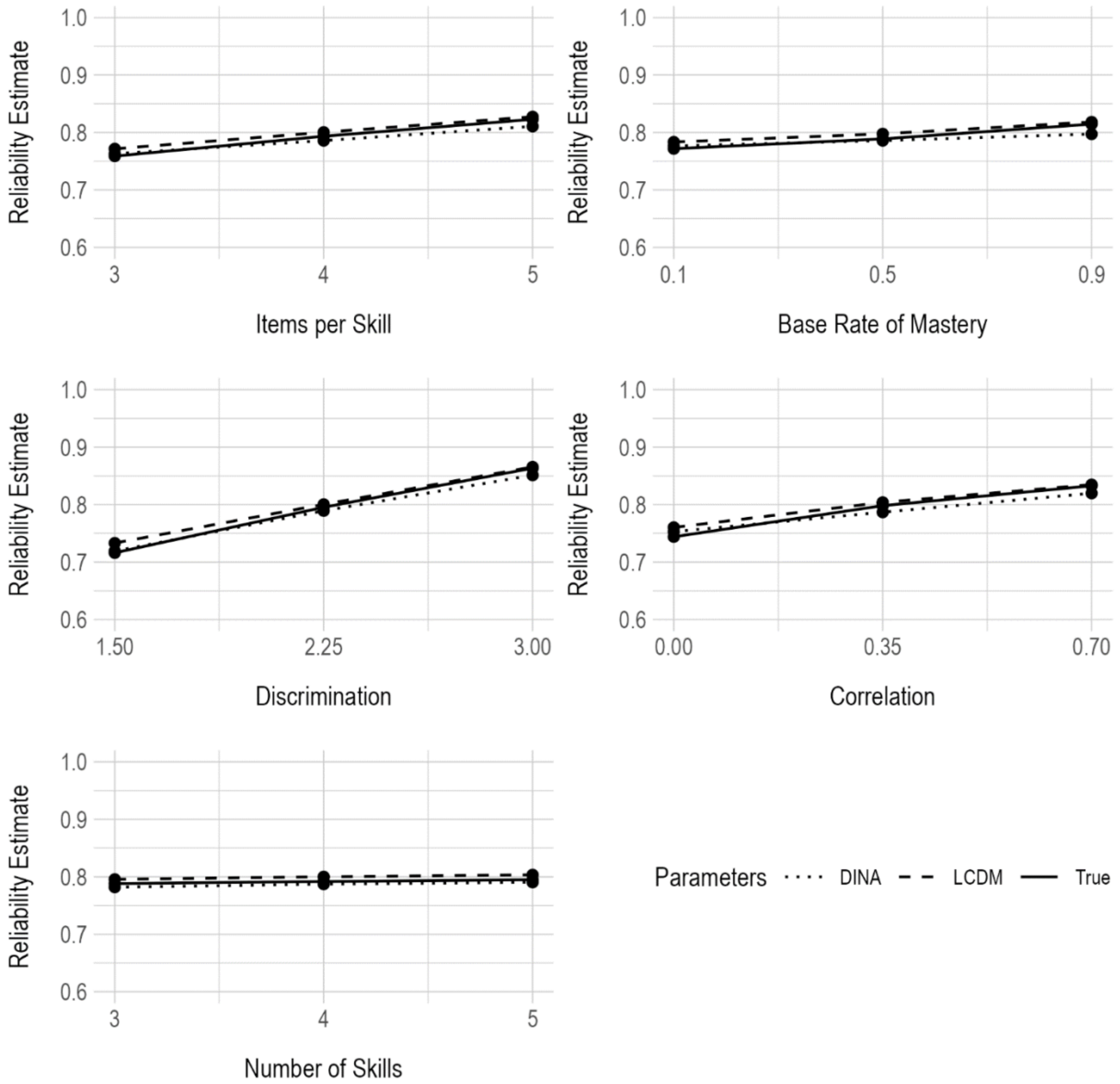
Model Fit

Figure 3 shows the average simulation-retest classification consistency for the first skill across each of the manipulated factors in the study and each set of item parameters (i.e., true, LCDM, and DINA). As expected, when using parameters from the DINA model that do not fit the true structure of the data, the reliability estimates are slightly lower than the estimates derived when using the true data-generating parameters or the LCDM estimates. The relatively small effect of misfit is likely an artifact of the Q-matrix generation. In the Q-matrix, each skill was always measured by two items in isolation (i.e., single-skill items) and one to three items that may or may not have measured a second skill (e.g., Table 2). For single-skill items, the LCDM and DINA models are equivalent (Rupp et al., 2010). Because the models are equivalent for single-skill items, misfit would be present only for the comparatively few numbers of items that measured multiple skills. As such, this study included only small to moderate levels of misfit, depending on how many items were simulated to measure multiple skills. Thus, it is likely that more items measuring multiple skills or items measuring more than two skills would increase the observed differences for the DINA model, as there would be a greater difference between the DINA model and the data-generating model.

In contrast to the results from the DINA model, the reliability estimates when using parameters from the LCDM that do fit the true structure of the data are slightly higher than the estimates derived when using the true data-generating parameters. This observation is especially true at the highest value of each of the study factors. The high values of these factors are typically associated with high quality assessments (e.g., highly discriminating items, longer test length, etc.). Across all simulation conditions, the average absolute difference between the simulation-retest classification consistency derived from the true and LCDM parameters was 0.0099, compared to a difference of 0.0168 between the true and DINA parameters.

Figure 3

Average Simulation-Retest Classification Consistency Across Study Factors, by Model

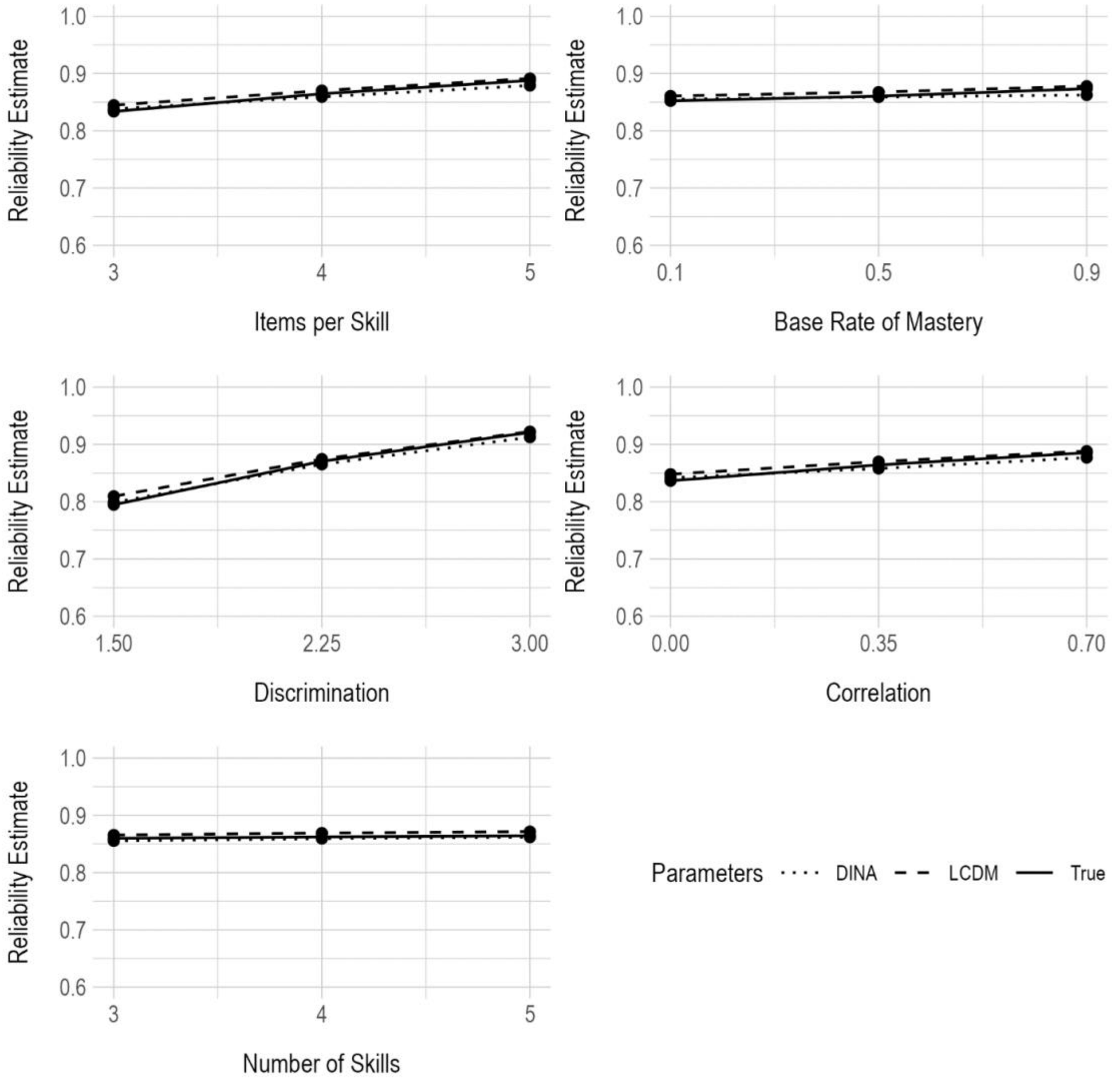


Classification accuracy shows a similar pattern in Figure 4. Again, we see that estimates of classification accuracy are generally slightly lower when the parameters come from a model that does not fit the data (i.e., the DINA model) and that estimates of classification accuracy are generally slightly higher when the parameters come from a model that does fit the data (i.e., the

LCDM). The differences are again most pronounced at the highest levels of each factor, but the differences are smaller overall than what we observed for the classification consistency. Across all simulation conditions, the average absolute difference between the true and LCDM estimates of classification accuracy was 0.0071, compared to an average absolute difference of 0.0110 between the true and DINA estimates (a difference of 0.004, compared to a difference of 0.007 for classification consistency).

Figure 4

Average Simulation-Retest Classification Accuracy Across Study Factors, by Model



Across all conditions, the estimates of classification accuracy were consistently higher than the estimates of classification consistency. Additionally, both the simulation-retest classification consistency and classification accuracy show patterns that are expected for a reliability metric. For example, both consistency and accuracy tend to increase as the number of items increases (top left of Figure 3 and Figure 4), as the items better differentiate between mastery classes (middle left of Figure 3 and Figure 4) and the correlation between skills increases (bottom right of Figure 3 and Figure 4).

Empirical Data Analysis

To evaluate the performance of the simulation-retest reliability method in a real-data setting, we applied the method to the data set for the grammar subtest of the ECPE (Templin & Hoffman, 2013), as the nonsimulation-based reliability estimates have been reported for the ECPE (Sinharay & Johnson, 2019). The ECPE is an internationally administered assessment of the grammatical rules in English at the Proficient level of the Common European Framework of Reference for Languages. More specifically, the ECPE assesses morphosyntactic rules, cohesive rules, and lexical rules. The ECPE is intended for secondary-school students and adults. The ECPE data set is available from the CDM package in R (Robitzsch et al., 2020), and it has previously been used to demonstrate the application of the nonsimulation-based classification-based reliability estimates for DCMs (Sinharay & Johnson, 2019). The data for the grammar subtest of the ECPE include 2,922 examinees and 28 items, with 13 items measuring the morphosyntactic rules, six items measuring cohesive rules, and 18 items measuring the lexical rules. The Q-matrix and the estimated structural and item parameters for the grammar subtest of the ECPE are available in Templin and Hoffman (2013).

In this empirical data analysis, we fit an LCDM to the ECPE data, and then we simulated 100 retests for each examinee using the estimated parameters from the LCDM. Given the many applications of this LCDM to this data set (e.g., Chen et al., 2018; Liu & Johnson, 2019; Templin & Bradshaw, 2014; Templin & Hoffman, 2013), we expect the parameter estimates to provide good model fit. We then estimated the simulation-retest estimates of classification consistency and accuracy as described previously and compared the simulation-retest estimates to the estimates of $\hat{P}_{ck}$ and $\hat{\tau}_k$ (denoted as $\hat{P}_{ak}$) by Johnson and Sinharay (2018; their Table 6 and Table 7). The $\hat{P}_{ck}$ and $\hat{\tau}_k$ are the previously described classification consistency and accuracy estimates defined by Johnson and Sinharay (2018) and Wang et al. (2015), respectively.

Empirical Data-Analysis Results

The reliability estimates from both the simulation-retest and nonsimulation-based methods are presented in Table 3. The simulation-retest reliability estimates were similar to the nonsimulation-based reliability estimates reported by Johnson and Sinharay (2018). The simulation-retest classification accuracy estimates were within .01 of the nonsimulation-based classification accuracy estimates. This high degree of similarity is expected, given the high degree of consistency between these measures that was observed in the simulation study, and indicates that the similarity between the two methods persists for real data. However, it is also worth noting that the simulation-retest reliability estimates were equal to or marginally larger than the nonsimulation-based reliability estimates. This marginal inflation, although never greater than .01, may indicate a slight tendency for the simulation-retest method to overestimate reliability.

Table 3

Comparison of Simulation-Retest and Nonsimulation-Based Skill-Level Reliability Estimates for the Data From the Examination for the Certificate of Proficiency in English

Measure	$\hat{P}_{ck}$	Simulation-retest consistency	$\hat{P}_{ak}$	Simulation-retest accuracy
Skill 1	.83	.85	.90	.92
Skill 2	.81	.82	.86	.86
Skill 3	.86	.87	.92	.93

Discussion

In this study, we compared the performance of a simulation-retest reliability method to nonsimulation-based methods that have previously been described in the literature. Although the simulated-retests method has been described and implemented in previous research (e.g., Thompson et al., 2019), additional research was needed to fully evaluate the estimates derived from such a method.

The findings from this paper demonstrate that simulated retests provide high-fidelity measures of classification consistency and classification accuracy for diagnostic assessments. When comparing the scores from simulated retests to the scores from an original data set, the simulated-retests method provided estimates of classification consistency and accuracy that were highly consistent with more traditional, nonsimulation-based methods. This similarity in the reliability estimates was true across all conditions evaluated in this study. Additionally, we demonstrated that the simulated-retest method demonstrates the expected properties of a reliability metric, such as increased reliability with longer assessments, more-discriminating items, and association between the measured constructs (de la Torre & Patz, 2005; DeVellis, 2006). Finally, an empirical analysis of ECPE data further demonstrated that the simulated-retests method produced reliability estimates similar to those of the nonsimulation-based reliability methods.

Best practice in the literature indicates reliability evidence should be presented at the same level at which the results are reported (e.g., Standards 2.2, 2.3, and 2.5, AERA et al., 2014; Sinharay & Haberman, 2009). For DCMs, when results are reported at the skill level, the reliability evidence should also be reported at the skill level. This guiding principle has motivated much of the existing research on reliability in DCMs (e.g., Cui et al., 2012; Johnson & Sinharay, 2018; Wang et al., 2015), where classification-based reliability at the skill and profile levels has been emphasized.

However, a limitation of existing classification-based reliability approaches is that they do not readily scale to other levels of reporting. For example, in addition to reporting the skill-level results, a testing program may also report results as the total number of skills mastered, a performance level for state accountability systems, or a pass/fail decision for certification and licensure. Consequently, it is important that reliability evidence can be reported at these levels.

This paper expands on previous work (e.g., Roussos et al., 2007; Thompson et al., 2019) to examine how simulation-retest estimates of classification accuracy and consistency compare to other methods. Because findings were generally consistent with other methods, we argue that simulated retests may be preferred because they can estimate reliability at multiple levels of reporting, not just the skill level (Thompson et al., 2019). As operational programs continue to adopt DCM-based assessments, the capacity to report results and provide reliability evidence at levels beyond just the skill level is important for meeting the needs of stakeholders.

We recognize that the simulated-retests method may not be necessary or preferable in all contexts. The process of simulating retests, calculating results for each retest, and summarizing the results with an appropriate agreement metric requires more time and computing than other reliability metrics for DCMs that provide equally useful information. However, when an assessment reports results at an aggregated level (e.g., an overall performance level), the simulated-retests method provides a consistent approach that can be used to report reliability for all levels of reported results. Thus, this method is an important tool for operational programs or accountability assessments that aggregate respondent-mastery results in addition to reporting individual skill-mastery statuses.

Because the simulated retests use the estimated model parameters to simulate the retests, model fit is a key component of the method. The results of the current study demonstrated that even the small to moderate amounts of misfit introduced in this study by using the DINA model may introduce bias in the reliability estimates. Therefore, practitioners implementing this method should carefully evaluate the fit of their model before using simulated retests to estimate reliability. Future work may consider the impact of different types and amounts of model misfit on the reliability estimates produced by simulated retests.

Conclusions

This study has positive implications for operational testing programs administering assessments that are scaled using DCMs. The simulation-retest reliability method allows for reliability evidence to be reported at multiple levels, which can support programs reporting both skill-mastery profiles and overall performance-level results. The current study demonstrates that the simulated-retests method generates reliability estimates that are consistent with nonsimulation-based methods and with the true reliability. These findings indicate that the simulation-retest reliability method produces accurate and consistent reliability estimates under a variety conditions. These findings are promising given the usefulness of the simulation-retest reliability method to operational testing programs in reporting reliability evidence to support the use of their assessments.

References

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

Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 297–327). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch13

Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice*, *38*(2), 79–88. https://doi.org/10.1111/emip.12247

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x

Chen, F., Liu, Y., Xin, T., & Cui, Y. (2018). Applying the M_2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in Psychology*, *9*, 1875. https://doi.org/10.3389/fpsyg.2018.01875

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2011.00158.x

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Patz, R. J. (2005). Make the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statisitcs*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), S50–S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 732–764. https://doi.org/10.1080/01621459.1954.10501231

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

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

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196

Johnson, M. S., & Sinharay, S. (2020). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*, *45*(1), 5–31. https://doi.org/10.3102/1076998619864550

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Liu, X., & Johnson, M. S. (2019). Estimating CDMs using MCMC. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 629–646). Springer Nature. https://doi.org/10.1007/978-3-030-05584-4_31

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *195*, 1–47. https://doi.org/10.1098/rsta.1900.0022

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2020). *CDM: Cognitive diagnosis modeling* (Version 8.1-12) [Computer software]. https://CRAN.R-project.org/package=CDM

Roussos, L. A., Dibello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive* *diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge University Press. https://doi.org/10.1017/CBO9780511611186.010

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature reivew and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement*, *7*(1), 46–49. https://doi.org/10.1080/15366360802715486

Sinharay, S., & Johnson, M. S. (2019). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 359–377). Springer Nature. https://doi.org/10.1007/978-3-030-05584-4_17

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339. https://doi.org/10.1007/s11336-013-9362-0

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010

Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*, *32*(4), 298–309. https://doi.org/10.1080/08957347.2019.1660345

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457–476. https://doi.org/10.1111/jedm.12096

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Records (1896–1970)*, *62*(40), 1432–1449. https://doi.org/10.2307/4586294

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3