Using Simulation to Evaluate Retest Reliability of Diagnostic Assessment Results

Brooke Nash, Amy K. Clark, W. Jake Thompson

University of Kansas

Overview

- Background
- Diagnostic Classification Models
  - Measuring reliability
  - Simulation-based retest reliability as an alternative
- Methods
- Example
- Discussion

Background

- If a test is administered twice and provides accurate measurement of knowledge, skills, and ability, the student 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 results are conflated with measurement error, which has an obvious negative impact on the validity of inferences made from the results.

Background (cont.)

- 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.

Purpose

- The purpose of this paper is to contribute to the conceptual understanding of simulation-based retest reliability by providing an overview of procedures and results from its application in an operational large-scale diagnostic assessment program.

Selecting a Reliability Method

- Depends on several factors, including the design of the assessment, the scoring model used to provide results, and availability of data.
- The guidelines put forth by the *Standards for Educational and Psychological Testing* specify a number of considerations for reporting reliability of assessment results.
  - Standard 2.2
  - Standard 2.5

Selecting a Reliability Method (cont.)

- While methods of obtaining “traditional” reliability estimates are well understood and documented, there is far less research on methods for calculating the reliability of results derived from less commonly applied statistical models, namely, diagnostic classification models (DCMs).

Diagnostic Classification Models

- DCMs are confirmatory latent class models that represent the relationship of observed item responses to a set of categorical latent variables.
- Whereas traditional psychometric models (e.g., IRT) model a single, continuous latent variable, DCMs model student mastery on multiple latent variables or skills of interest.
  - 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 skill level.

Model Probabilities

- Based on the collected item response data, the model determines the overall probability of students being classified into each latent class for each skill.
  - The latent classes for DCMs are typically binary mastery status (master or nonmaster).
- This base-rate probability of mastery is then related to students' individual response data to determine the posterior probability of mastery.
- The posterior probability is on a scale of 0 to 1 and represents the certainty the student has mastered each skill.

Interpreting the Posterior Probability

- Values closer to extremes of 0 or 1 indicate greater certainty in the classification.
  - 0 indicates the student has definitely not mastered the skill
  - 1 indicates the student has definitely mastered the skill
- In contrast, values closer to 0.5 represent maximum uncertainty in the classification.
- Results for DCMs may be reported as the mastery probability values or as dichotomous mastery statuses when a threshold for demonstrating mastery is imposed (e.g., .8).

Reliability of DCMs

- The DCM scoring approach is unique in that the probability of mastery provides an indication of error, or conversely confidence, for each skill and examinee.
- However, it does not provide information about consistency of measurement for the skill or assessment as a whole.

Reliability of DCMs (cont.)

- Traditional approaches to reliability are not appropriate and alternate methods must be considered for reporting the reliability of DCM results.
- “Standard reliability coefficients, as estimated for assessments modeled with a continuous unidimensional latent trait, do not translate directly to discrete latent space modeled cognitive diagnostic tests” (Roussos et al., 2007).

Other Considerations

- Test design and the extent to which the assumptions about the assessment are met
  - For instance, the Cronbach’s coefficient alpha assumes tau-equivalent items (i.e., items with equal information about the trait but not necessarily equal variances), though not all assessments are designed to meet this assumption.
- Consistency with level at which results are reported
  - Sinharay & Haberman (2009) argued that, to support the validity of inferences made from diagnostic assessments reporting mastery at the skill level, reliability must be reported at the same level.

DCM Reliability Indices

- Researchers have begun developing reliability indices that are more consistent with diagnostic scoring models.
  - A modified coefficient alpha was calculated for an attribute hierarchy model using existing large-scale assessment data (Gierl, Cui, & Zhou, 2009).
    - Used IRT ability estimates for calibration and scoring, rather than an attribute-based scoring model
  - The cognitive diagnostic modeling information index (Henson & Douglas, 2005) reports reliability using the average Kullback-Leibler distance between pairs of attribute patterns.
    - Does not report reliability for each attribute itself
- For operational assessments that are calibrated and scored using a diagnostic model and report performance via individual skill mastery information, alternative methods for reporting reliability must be explored.

Simulation-Based Retest Reliability

- In light of these concerns, simulation-based methodology has emerged as a possible solution for reporting reliability of diagnostic assessment results.
- Conceptually, a simulated second administration of an assessment can provide a means for evaluating retest reliability in the traditional sense (i.e., consistency of scores across multiple administrations).

Interpretation

- While the simulation-based approach differs from traditional methods and instead reports the correspondence between true and estimated mastery statuses, the interpretation of the reliability results remains the same.
  - Values 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 student differences on the construct measured by the assessment.

Benefits

- Using real-data collection approaches, second test administrations are susceptible to several additional construct irrelevant sources of error (e.g., learning, forgetting, practice).
  - Conversely, simulated second administrations that are based on real student data and calibrated model parameters closely mimic real student response patterns sans human error.
- Finally, as attempts to conduct a second administration of an assessment are usually met with concerns related to policy, cost, time, resources and overall feasibility, simulating a theoretical second administration becomes a particularly valuable alternative.

METHODS FOR SIMULATION-BASED RELIABILITY

General Approach

- Generate a second set of student responses based on actual student performance and calibrated-model parameters; score real test data and simulated test data; and compare estimated student results with the results that are true from the simulation.

Skill Mastery

- In the context of using DCM to calibrate and score the assessment, student performance is the set of mastery statuses for each skill.
- Mastery status is determined based on a specified threshold to distinguish masters and non-masters, again.
  - In applications of this methodology, the threshold value may vary depending on the design of the assessment, student population, stakeholder feedback, or other factors.

Steps in Simulation

1. **Draw student record**. Draw with replacement a student record from the operational dataset. The student’s mastery statuses from the operational scoring for each measured skill serve as the true values for the simulated student.
2. **Simulate second administration.** For each item the student was administered, simulate a new response based on the model-calibrated parameters, conditional on 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 students.

Calculating Reliability

- Estimated skill mastery statuses are compared to the known values from the simulation. Reliability results are calculated based on the 2x2 contingency table of estimated and true mastery status for each measured skill.

		Estimated	
		Master	Non-Master
True	Master	$p \times p$	$p(1-p)$
	Non-Master	$(1-p)p$	$(1-p)(1-p)$

Reliability Indices

- Three metrics of association between the true and estimated mastery status for each skill assessed are described:
  - Tetrachoric correlation between true and estimated mastery status
  - Correct classification rate and the chance-corrected correct classification Cohen's Kappa for the mastery status of each skill
  - Pearson correlation between true and estimated number of total skills mastered within the subject

Reporting Reliability

- The inclusion of multiple metrics of association in technical documentation provides a fuller picture of the reliability of the assessment than any one metric can provide.
- Results for each skill can be summarized in tabular form by subject, grade or other level of reporting.
- Depending on the number of skills measured, it may be necessary to report aggregated results rather than reporting reliability on individual skills.

SIMULATION-BASED RELIABILITY EXAMPLE

Dynamic Learning Maps (DLM) Alternate Assessment System

- The DLM System administers assessments to approximately 90,000 students annually in a 17-state consortium.
- Assessments are available in grades 3-8 and high school in English language arts, mathematics, and science.
- Each alternate content standard is measured at multiple levels of cognitive skill complexity (known as linkage levels), with each varying in complexity from the grade-level target skill.

Scoring & Reporting of Results

- The DLM diagnostic assessment system was built from a set of underlying learning map models.
  - Each linkage level (skill) measures one or more nodes in the learning map model; linkage levels are the basis for reporting results of the assessment.
- Assessment results are calibrated and scored using a latent class DCM to produce student mastery profiles, summarizing mastered skills for each content standard.
- Results are reported at multiple levels including at the skill level (within each content standard), within larger content strands, and for the overall subject area.

Setting a Threshold for Mastery

- The DCM used to calibrate and score DLM assessments produces student-level posterior probabilities for each skill for which a student was assessed.
- A threshold was established to make mastery status classifications based on the probabilities for each skill.
  - The standard setting process was based on a combination of analysis of impact data and stakeholder feedback, which included both the consortium governance board and Technical Advisory Committee.
  - This process resulted in a mastery threshold of .8

Procedures

- Data from the 2017 operational administration of the DLM assessments were used to simulate student response data as the second administration for evaluating retest reliability.
- The number of replications was set to 2,000,000 for each subject (English language arts, mathematics and science) to ensure adequate sample size when calculating reliability.
- Steps followed those previously outlined

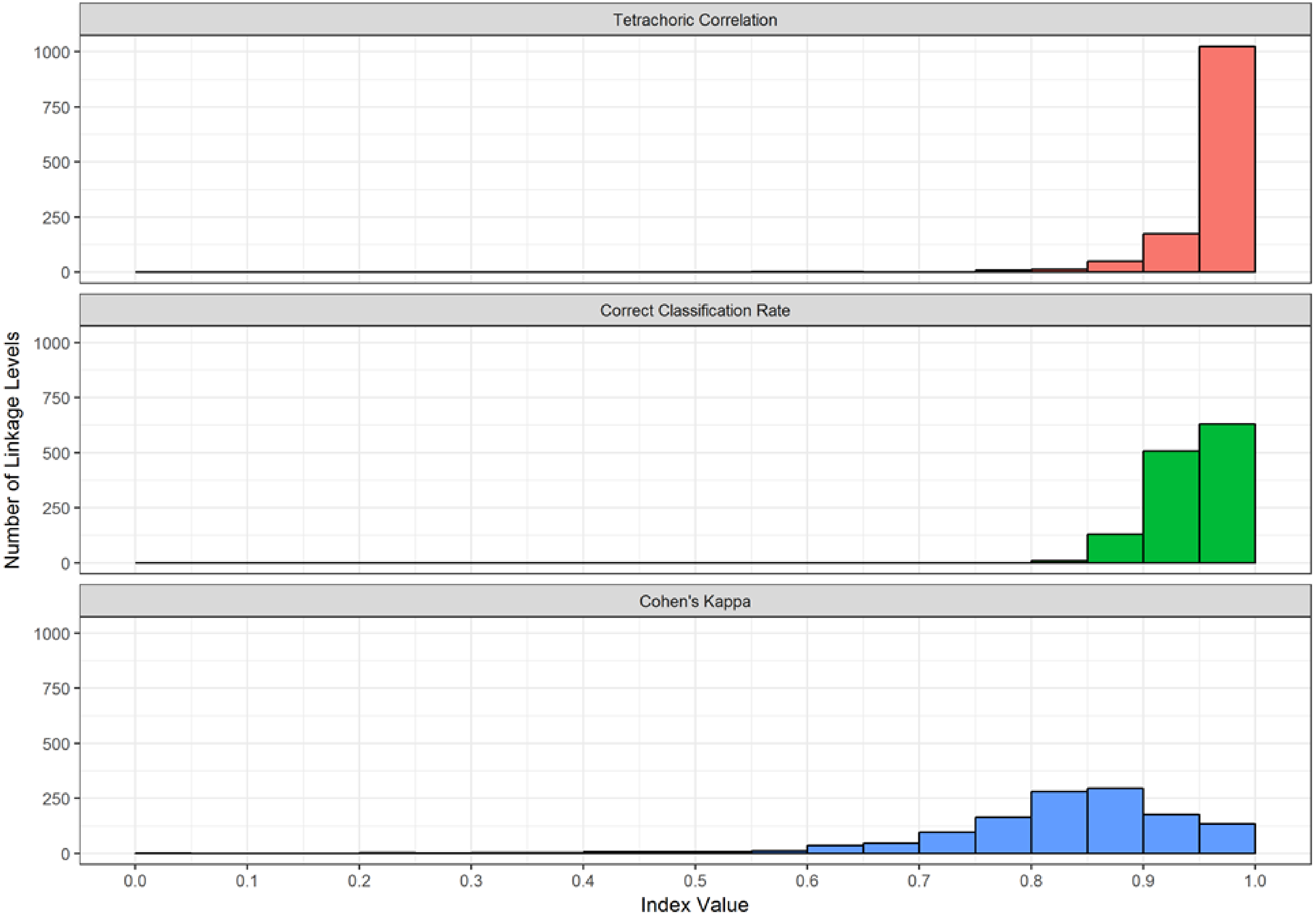
Results

- Reliability estimates calculated for a total of 1,410 skills measured across all grades and subjects
- Indices included tetrachoric correlations between true and estimated mastery statutes, correct classification rates for the mastery status of each skill, and the chance-corrected correct classification Cohen’s Kappa for the mastery status of each skill.
  - While example reliability results provided here are at the skill level, mastery statuses of skills can also be aggregated to other levels of reporting, for example, at the subject level (see Thompson, Clark & Nash, 2018).

Proportion of Skills Falling within a Specified Index Range

Reliability Index	Index Range								
	< .60	.60–.64	.65–.69	.70–.74	.75–.79	.80–.84	.85–.89	.90–.94	.95–1.0
Tetrachoric Correlation	0.004	0.001	0.002	0.002	0.002	0.010	0.017	0.096	0.866
Correct Classification Rate	0.000	0.000	0.000	0.001	0.002	0.006	0.058	0.330	0.603
Kappa	0.038	0.016	0.021	0.057	0.104	0.177	0.221	0.181	0.184

Example Summary of Reliability Results



Summary of Results

- Summaries indicate that, in general, the skills measured by the assessment show strong evidence of consistency of measurement across administrations.
- Because of the high threshold for skill mastery for DLM assessments (0.8), results such as these are expected and reflect, in part, the consistency of classifying students as masters or nonmasters inherently built into diagnostic mastery decisions themselves.
- Moreover, the results reflect an upper bounds estimate of reliability to the extent the data fit the model.

DISCUSSION

DCM Reliability

- As diagnostic assessments become more prevalent, alternatives to traditional reliability methods must be explored.
- Reliability estimates from DCMs are expected to be higher than those using traditional scoring models for a couple of reasons.
  - First, as mentioned, the mastery threshold itself can create highly replicable results.
  - Second, the goal of DCMs is to assign a classification status on one or more categorical latent traits. Thus, the coarser level of measurement in DCMs (typically master or nonmaster) results in a more precise classification decision than continuous latent trait analogues (Templin & Bradshaw, 2013).

Simulation-Based Retest Reliability

- While the current application of the simulation-based reliability methodology was for a diagnostic assessment that utilizes DCM, the concept of a simulated second administration of an assessment as a method for collecting retest data may be applied to other scoring models.
- As the collection of real retest data is often infeasible and is susceptible to measurement error that can be attributed to the data collection design, simulating retest data is a worthwhile alternative to consider.

Future Research

- Additional research is needed to further evaluate its use.
  - A simulation study could be conducted where the reliability of the assessment is known and compared to the reliability estimates calculated from the simulated data when mastery threshold and item parameters are varied.
  - Similarly, given that the simulation method assumes perfect model fit, which is not possible in application, another informative study would be to introduce varying levels of model misfit and evaluate the impact on reliability estimates.

Thank You!

- Correspondence concerning this paper should be addressed to Brooke Nash, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045; 785-864-8191; bnash@ku.edu.