Using Simulation to Evaluate Retest Reliability of Diagnostic Assessment Results

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

University of Kansas

Author Note

Paper presented at the 2018 annual meeting of the National Council of Measurement in Education, New York, NY. 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.

This work was partially supported by the U.S. Department of Education, Office of Special Education Programs under Grant 84.373 100001. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

Acknowledgment: The authors wish to acknowledge Dr. Jonathan Templin for his contribution to the design and analysis of the reliability methodology for the Dynamic Learning Maps Assessments.

Abstract

As diagnostic assessment systems become more prevalent as large-scale operational assessments, consideration must be given to the method of reporting reliability. Alternatives to traditional reliability methods must be explored that are consistent with the design, scoring, and reporting level of diagnostic assessment systems. One method for evaluating retest reliability when practical constraints make a second empirical administration infeasible is with the use of simulation methodology. The purpose of this paper is to summarize the method and application of using a simulated second test administration to report reliability for one large-scale operational diagnostic assessment program. Using operational administration data, student response data was simulated based on model-calibrated parameters. Reliability estimates were calculated to provide a measure of association between true and estimated mastery of skills. Overall, results provide support for reporting reliability via simulation-based methods and for the valid interpretation and use of skill mastery information provided in diagnostic score reports. This paper includes a summary of the methods used, presentation of example results, broad implications for its application within the measurement field and future directions.

Keywords: reliability, simulation, diagnostic testing, assessment, score reporting

Using Simulation to Evaluate Retest Reliability of Diagnostic Assessment Results

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 often-given example, using a measuring tape to measure the length of a box should produce the same results each time. The results should be highly consistent from one measurement to the next. The same can be said of measurement in education. 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.

However, 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 coefficient alpha (Cronbach, 1951) 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 each half of the assessment as separate forms administered at the same time.

Selection of a method for evaluating reliability of an assessment 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* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014) specify a number of considerations for reporting reliability of assessment results. Standard 2.2 indicates, "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" (AERA et al., 2014, p. 42). Further, Standard 2.5 indicates "Reliability estimation procedures should be consistent with the structure of the test" (AERA et al., 2014, 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 also dominated the reliability literature (e.g., Haertel, 2006; Traub & Rawley, 1991). 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, also known as cognitive diagnosis models (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, 2017; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). 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.

To provide detailed profiles of student 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, 1995). 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 (i.e., the structural parameter) 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. Values closer to extremes of 0 or 1 indicate greater certainty in the classification, whereby a value of 0 indicates the student has definitely not mastered the skill, and a value of 1 indicates the student has definitely mastered the skill. In contrast, values closer to 0.5 represent maximum uncertainty in the classification. A mastery probability of 0.5 indicates the model cannot distinguish whether the student has mastered the skill based on the available response data; the student is just as likely to be a master as a nonmaster. 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)

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. Furthermore, because assessment results are the collection of skill mastery results rather than a

raw or scaled score, traditional approaches to reliability are not appropriate and alternate methods must be considered for reporting the reliability of results.

Measuring Reliability of DCMs

Because DCMs are a fairly recent development in the measurement community, there is limited documentation as to how reliability should be reported for these assessments. As such, perhaps it is easier to start this section with a discussion on how reliability cannot be measured for DCMs. As pointed out by Roussos et al. (2007), "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". For example, in item response models the inverse of reliability, which is the standard error of measurement, is based on the calculation of Fisher's information, which involves differentiating the likelihood function with respect to the continuous latent trait. However, when the likelihood is not a smooth function, such is the case with categorical latent traits, the levels of the trait cannot be differentiated (e.g., Henson & Douglas, 2005; Templin & Bradshaw, 2013).

The selection of a reliability method also depends on the 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. Take, for instance, the case of an adaptive test that is designed to align test items to each examinee's ability level. Not only do examinees take different items, but those items may provide more or less information about the trait depending on the examinee's ability level. Similarly, diagnostic assessments would likely not meet the assumption of tau-equivalence

required by Cronbach's alpha (and any metrics it subsumes, such as Spearman-Brown) because they are intentionally designed to measure multiple latent traits.

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. They also noted that this was a critical aspect missing from many diagnostic assessment applications implemented up to the time of publication. Templin & Bradshaw (2013) surmised that the lack of reporting of reliability for DCMs is due to the lack of a well-defined concept of reliability for DCMs. As diagnostic assessment systems transition into being implemented as operational assessment programs, the reporting of reliability evidence is critical to the validity of inferences that can be made from results. Based on recommendations put forth by the *Standards* (AERA et al., 2014), as well as the commentary by Sinharay & Haberman (2009), it is critical that assessments scored with diagnostic modeling to report results at the fine-grained skill-mastery level must also provide reliability evidence at a commensurate level, obtained using a method consistent with the scoring procedure.

To this end, researchers have begun developing reliability indices that are more consistent with diagnostic scoring models. For example, a modified coefficient alpha was calculated for a retrofitted attribute hierarchy model using existing large-scale assessment data (Gierl, Cui, & Zhou, 2009). The modified alpha provides the ratio of true score variance to observed score variance for each individual attribute measured by a five-attribute model. However, the attribute hierarchy method makes use of IRT ability estimates for calibration and scoring, rather than an attribute-based scoring model, to assign examinees to the most likely profile of attribute mastery. Similarly, the cognitive diagnostic modeling information index (Henson & Douglas, 2005) reports reliability using the average Kullback-Leibler distance

between pairs of attribute patterns, rather than reporting 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). While the simulation-based approach differs from traditional methods (e.g., Coefficient Alpha), and instead reports the correspondence between true and estimated mastery statuses, the interpretation of the reliability results remains the same. That is, 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.

Roussos et al. (2007) explained how simulated data obtained from calibrated model parameters (based on real data) can be used to produce summary statistics for evaluating the model, including several types of reliability indices. Specifically, 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. Similarly, the proportion of times each examinee is classified correctly for each skill was also described as providing an estimate of the correspondence between true and estimated skill classification.

Templin & Bradshaw (2013) conducted a research study using simulation 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 of DCM involved retrofitting the model to existing large-scale assessment data designed to measure a single construct, and the assignment of items to attributes was imposed. The researchers used posterior probabilities of mastery to calculate the probability of being assigned to each mastery profile, and compared this to random draws from the theta distribution for the IRT-scored assessment. Reliability results comparing estimated probabilities of assignment to each possible mastery status across the first administration and hypothetical second administration were reported with a tetrachoric correlation for each attribute in the model. While their main findings demonstrated that DCM produced higher reliability estimates than those obtained from the IRT model for a same-length test, they also demonstrated that estimated second test administrations could be useful for evaluating reliability.

Simulating retest data can also be useful for reporting multiple metrics of reliability. As demonstrated in Roussos et al., (2007), by treating the results from the real-data calibration as true and the simulated results as estimated, correct classification rates can be calculated. Extending this concept, the frequency of correct classifications can be calculated and aggregated across examinees as a method for providing the correct classification rate by skill.

While the current study focuses on use of simulation-based reliability methodology in the context of DCMs, it is our belief that this methodology may be valuable in more traditional (i.e., IRT and CTT) contexts as well to provide a fuller description of consistency of measurement when collecting real retest data is impractical. In addition to conceptual consistency with

traditional methods, a simulation-based retest method has several other 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. In this sense, simulation-based methodology may remove additional sources of error that may be observed in real data retest approaches. 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.

As the use of DCM within the measurement field expands, and combines with the limited practicality of collecting retest data, the use of simulation should be further explored as a suitable alternative for reporting reliability of large-scale assessments. 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.

Dynamic Learning Maps Alternate Assessment System

The Dynamic Learning Maps (DLM) Alternate Assessment 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. The assessment measures student performance on alternate content standards. Each standard is measured at multiple linkage levels, with each varying in complexity from the grade-level target skill. In English language arts and mathematics, each standard is available for assessment at five linkage levels; in science each standard is available at three linkage levels.

The DLM diagnostic assessment system was built from a set of underlying learning map models. The test development process connects content standards to nodes in the map. 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, rather than a scale score for a single latent trait. 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. Because reliability should be reported consistent with the test structure and intended uses of results, as recommended by the *Standards* (AERA et al., 2014), reporting reliability with traditional methods is not appropriate for this assessment.

Methods for Simulation-Based Reliability

The general approach to a simulation-based reliability method is to 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. That is, once student response data has been collected, calibrated, and scored, a second administration can be simulated based on the known model parameters from the first administration. Student records are drawn from operational data to simulate a second administration based on the actual set of items each examinee has taken, which means that the two administrations are perfectly parallel.

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, recognizing the values further from .5 indicate greater certainty in the classification. In applications of this methodology, the threshold

value may vary depending on the design of the assessment, student 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. This is the level of reporting results for diagnostic assessments, and the level at which reliability must be summarized. Because the scoring model produces mastery decisions, the term *results* is used in place of *scores* throughout this paper.

The specific steps for a DCM-based simulation to produce a theoretical a second administration are as follows:

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

The simulation-based method used to report reliability results draws from the design of a diagnostic assessment system; therefore, reliability results are provided for each skill measured. To calculate reliability indices, the estimated skill mastery statuses are compared to the known

values from the simulation. Specifically, reliability results are calculated based on the 2x2 contingency table of estimated and true mastery status for each measured skill where the probability of mastery across two administrations (i.e., true and estimated) can be calculated as $p \times p$ and the probability of each mastery/non-mastery status for a single skill can be defined, as shown in Table 1. Results based on the contingency table can also be aggregated across skills to quantify the assessment's internal consistency and aid in the interpretation of results summarized in score reports.

[Insert Table 1 about here]

As with any contingency table, a number of summary statistics are possible for describing results. Three metrics of association between the true and estimated mastery status for each skill assessed are described here. Consistent with Templin & Bradshaw (2013), reliability results may be summarized with the tetrachoric correlation between true and estimated mastery status. Results can also reported as the correct classification rate for the mastery status of each skill and the chance-corrected correct classification Cohen's Kappa for the mastery status of each skill. Kappa values between 0.6 and 1.0 indicate substantial-to-perfect agreement between the true and estimated mastery status (Landis & Koch, 1977). Parallel to more traditional methods of reporting reliability for total or scaled scores, a Pearson correlation between true and estimated number of skills mastered within the subject could also be calculated.

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. Once calculated, reliability results for each skill can be summarized for technical documentation purposes 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

An example of the simulation-based reliability method is described for the DLM assessment system. 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 (Clark, Nash, Karvonen, & Kingston, 2017) for specifying a mastery threshold 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, which was selected due to stakeholder desire for the value to be far enough from the point of maximum uncertainty (.5), but also taking into consideration the sometimes variable performance of students with the most significant cognitive disabilities who take alternate assessments (see Dynamic Learning Maps Consortium, 2016 for more information on this process).

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.

Following the general procedures for simulating student response data for a theoretical second administration of the assessment, the DLM procedure began with drawing, with replacement, a student record from the 2017 operational dataset. The student's mastery statuses from the operational scoring for each measured skill served as the true values for the simulated

student. For each item the student was administered, a new response based on the model-calibrated parameters was simulated, conditional on mastery status for the skill. The model-calibrated parameters were the same as those used to score the 2017 operational assessments. Using the operational scoring method, the simulated responses were scored using the .8 threshold for mastery imposed on the posterior probability of mastery obtained from the model. For DLM assessments, additional scoring rules are included in the operational scoring model to prevent the model from being overly influential. The first is a percent correct scoring rule, whereby mastery status is obtained for students who respond to at least 80% of items measuring the skill correctly. The second is a "two-down" scoring rule, whereby mastery status is obtained for a skill two levels down in the learning map models from the lowest level assessed but not mastered. For more information about DLM scoring rules, please see Chapter V of the *2014–2015 Technical Manual – Year-End Model* (Dynamic Learning Maps® Consortium, 2016). As mentioned, these steps were repeated for 2,000,000 simulated students in each subject.

For DLM assessments, the simulation-based reliability method resulted in reliability estimates for a total of 1,410 skills measured across all grades and subjects. Reliability estimates for each skill were calculated using 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).

Because of the number of skills measured by DLM assessments, reliability evidence is summarized in technical documentation. An example summary of simulation-based reliability

results for DLM assessments is shown in tabular form in Table 2, and in graphical form, as shown in Figure 1.

[Insert Table 2 about here]

[Insert Figure 1 about here]

Across measures of association and subjects, the DLM assessment reliability 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. Had a lower threshold been chosen, reliability results would similarly reflect reduced consistency in classifying students as masters or nonmasters. Moreover, the results reflect an upper bounds estimate of reliability to the extent the data fit the model.

Discussion

As diagnostic assessments become more prevalent as an alternative to IRT and classical test theory methods for calibration and scoring, alternatives to traditional reliability methods must be explored. This study summarizes a simulation-based method and provides an example for one diagnostic assessment system, whereby reliability results were summarized at the level of reporting (skills) for nine grades and three subjects. Overall, reliability evidence obtained from the simulation methodology for DLM assessments indicates a high-degree of consistency of measurement. These results were expected for several reasons. First, as mentioned, the mastery threshold applied to the posterior distribution to determine mastery status necessarily results in a high degree of certainty in mastery decisions. In other words, the threshold itself created highly replicable results. A second related reason that the results were expected is due to the nature of

DCMs. While the goal of traditional models is to locate the point on a continuous scale that best describes the amount of the trait that a student's possesses, 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).

Because diagnostic assessments produce fine-grained, highly-actionable score reports, the evaluation of reliability is critical to score report interpretation and the utility of reports to inform instructional decision-making. When there is less variability in results, and reliability is high, greater confidence can be placed in score report results because there is less measurement error included in the calculation of the results. This has important implications for teachers using score reports from diagnostic assessments to determine next steps for instruction, instructional groupings, and planning individualized instructional trajectories (e.g., individualized education plans). As such it is imperative that reliability methods yield accurate representations of the consistency of measurement for diagnostic assessments and the skill mastery information reported.

Furthermore, 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 can be applied to other scoring models, such as CTT and IRT. 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 for calculating the reliability of assessments that use any scoring model.

Considerations

As with any study, limitations are observed with the simulation-based reliability method. Because the simulation-based reliability method relies on the calibrated model parameters, evidence of model-data fit is imperative for supporting the overall validity of inferences that can be made from the results, as well as the utility of the reliability metric to summarize consistency of measurement. Therefore, the reliability estimates are considered upper bounds, and are reliant on the extent to which the model fits the data.

Additionally, the use of a simulation-based method for calculating reliability is more computationally-intensive than traditional methods. The analyses presented in this paper were conducted on a high performance computing platform; however, it should be noted that the amount of computing resources necessary is highly dependent on the assessment design, including number of grades, subjects, and test blueprints. In any case, when compared to the cost, time, and resources needed to conduct retest studies with real students, the computational burden of the simulation method may be overall less resource dependent than a real data method, but may require additional computing resources.

Future Research

While this paper provides a conceptual framework for and operational application of a simulation-based methodology for calculating retest reliability, additional research is needed to further evaluate its use. For example, 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.

Given that the highly consistent results found in this study are known to be a function of both the dichotomous mastery decision being made about each skill and the moderately high threshold applied to make that decision, it is difficult to discern the degree to which the simulated retest data (i.e., true mastery statuses) resembled the operational data as a result of the simulation procedure itself. In other words, did the simulation procedure produce retest data that resembled real student data on perfectly parallel forms to the greatest extent possible (minus measurement error associated with time-related factors)? Conversely, did the simulation procedure produce retest data that essentially mirrored the operational data due to the specifications of the procedure? Additional analyses should be conducted to evaluate the consistency of the means from the posterior distributions (i.e., rather than the classification decision) for the estimated and true mastery probabilities. For example, scatterplots of true and estimated mastery probabilities, particularly between the 0.4 and 0.6 range could be used to evaluate the consistency of the model-based probabilities.

Overall, the methods and outcomes summarized here provide support for simulation-based reliability metrics for reporting consistency of measurement for diagnostic assessment systems. As use of these systems expands, additional research should be conducted to evaluate its application across assessment systems of various complexity.

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Backhaus, J., Junghanns, K., Broocks, A., Riemann, D., & Hohagen, F. (2002). Test-retest reliability and validity of the Pittsburgh Sleep Quality Index in primary insomnia. *Journal of Psychometric Research, 53,* 737–740.

Bradshaw, L. (2017). Diagnostic classification models. In A. Rupp & J. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp.297–327). Malden, MA: Wiley.

Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice, 36*(4), 5–15. doi:10.1111/emip.12162

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. doi: 10.1177/0146621612445470

Dynamic Learning Maps Consortium. (2016). *2014–2015 Technical Manual – Integrated Model.* Lawrence: University of Kansas, Center for Educational Testing and Evaluation.

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology, 28*, 1095–1112.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 293–313.

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

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Washington, DC: American Council on Education/Praeger.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262–277. doi: 10.1177/0146621604272623

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363–374. doi: 10.2307/2529786

Littleton, A. C., Register-Mihalik, J. K., & Guskiewicz, K. M. (2015). Test-retest reliability of a computerized concussion test: CNS vital signs. *Sports Health, 7,* 443–447. doi: 10.1177/1941738115586997

March, J. S., & Sullivan, K. S. (1999). Test-retest reliability of the Multidimensional Anxiety Scale for Children. *Journal of Anxiety Disorders, 13,* 349–358.

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology, 56,* 730–735.

Nakayama, Y., Covassin, T., Schatz, P., Nogle, S., & Kovan, J. (2014). Examination of the test-retest reliability of a computerized neurocognitive test battery. *The American Journal of Sports Medicine, 42,* 2000–2005. doi: 10.1177/0363546514535901

Roussos, L. A., DiBello, L. V., Stout, W., Hartz S. M., Henson, R. A., & Templin, J. (2007). The Fusion Model skills diagnosis system. In J. Leighton, & M. Gierl (Eds.), *Cognitive*

diagnostic assessment for education: Theory and applications. New York: Cambridge University Press.

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

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

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chapman, R.L. Brennan (Eds.) *Cognitively diagnostic assessment* (pp. 327 – 359). Hillsdale NJ: Lawrence Erlbaum Associates.

Dynamic Learning Maps Consortium. (2016b). *2014-2015 Technical Manual – Year-End Model*. Lawrence, KS: University of Kansas.

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

Thompson, W.J., Clark, A.K. & Nash, B. (2018). *Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting*. Presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*(1), 37-45. doi: 10.1111/j.1745-3992.1991.tb00183.x

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California Verbal Learning Test–second edition: Test-retest reliability, practice effects, and reliable

change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology, 21*, 413–420.

Table 1

Contingency Table for True and Estimated Mastery Status from Reliability Simulation for Single Skill Measured by the Assessment

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

Table 2

Example Summary of Reliability Results for Skills (Linkage Levels) Measured by the DLM Alternate Assessment: 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

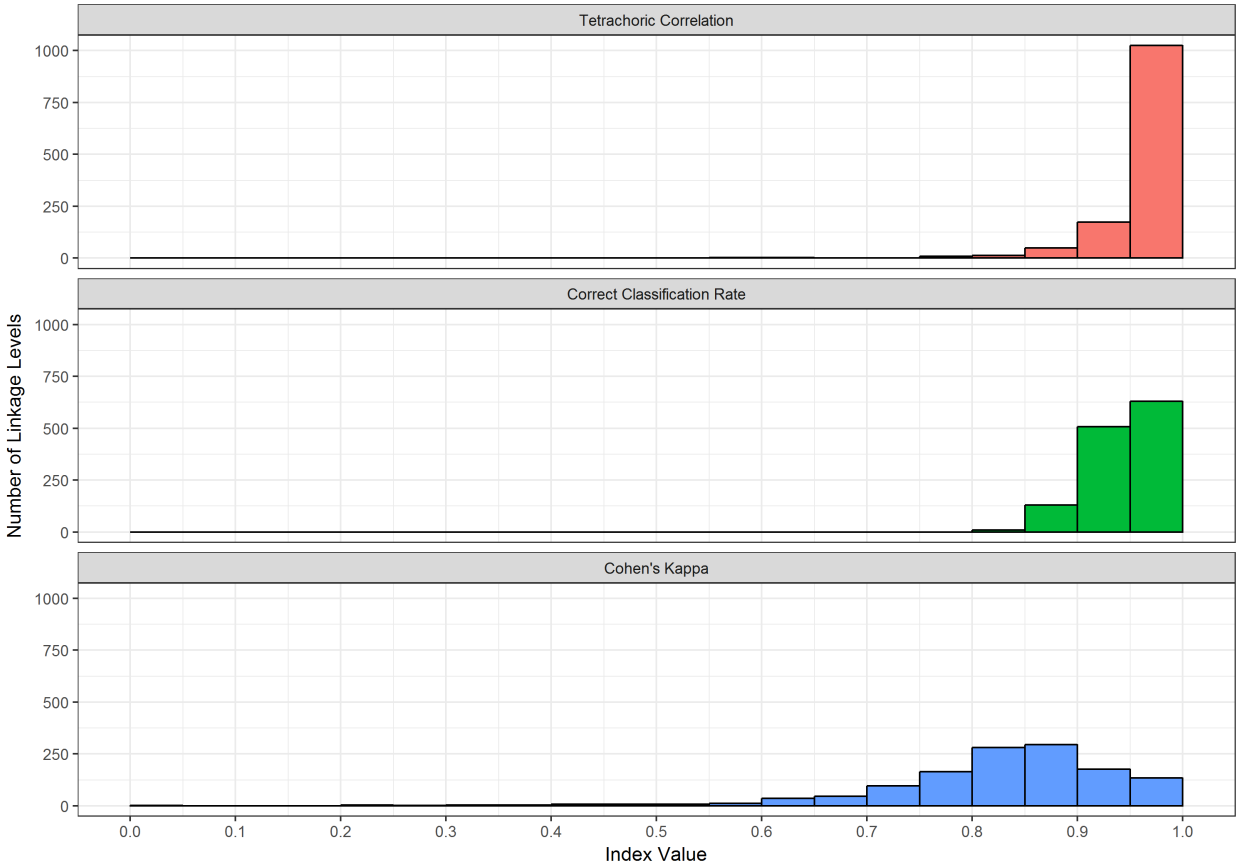

Figure 1. *Example summary of reliability results for skills (linkage levels) measured by the DLM alternate assessment.*