Abstract

As the use of diagnostic assessment systems transitions from research applications to large-scale assessments for accountability purposes, reliability methods that provide evidence at each level of reporting are needed. The purpose of this paper is to summarize one simulation-based method for estimating and reporting reliability for an operational, large-scale, diagnostic assessment system. This assessment system reports the results and associated reliability evidence at the individual skill level for each academic content standard and broader content strands. The system also summarizes results for the overall subject using achievement levels, which are often included in state accountability metrics. Results are summarized as measures of association between true and estimated mastery status for each level of reporting.

Keywords: diagnostic assessment, reliability, reporting, test–retest

Measuring Reliability of Diagnostic Mastery Classifications at Multiple Levels of Reporting

For assessment to be meaningful, results should provide stakeholders not only high-level evidence of student performance in the assessed constructs but also sufficiently fine-grained information to guide actionable next steps in instruction and learning. Assessment systems that provide diagnostic feedback are the object of increased academic and operational attention (e.g., Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010) because they can provide fine-grained information about what students know and can do. Instead of reporting a single score value on a broad construct of interest (e.g., mathematics, reading), diagnostic assessments can provide information about student mastery of many discrete skills or attributes measured by the assessment. Teachers, parents, and students can then use specific skill-mastery information to determine meaningful next steps for teaching and learning activities.

In addition to the fine-grained reporting that makes diagnostic assessments useful tools in instruction, many state accountability models require aggregated results in the form of achievement levels. Therefore, diagnostic systems that are also used for accountability reporting purposes may require multiple levels of reporting, such as by skill, content strands, and overall performance in the subject.

Regardless of the grain size of reported results, a critical aspect of any assessment system is the precision of the reported results. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 2014) state that reliability should be provided "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). For diagnostic assessment systems, the grain size at which student results are reported, as well as the assessment system's unique design and scoring method should be considered.

Whereas classical test theory (CTT; see Lord, Novick, & Birnbaum, 1968) and item response theory (IRT; see Hambleton, 1991) are based on continuous ability estimates (the raw score and a latent ability measure, respectively), diagnostic assessments use a categorical ability estimate. Because of the unique scoring and reporting considerations associated with diagnostic assessments, traditional approaches to reporting reliability must be modified accordingly. The authors of this paper detail methods for reporting reliability of diagnostic assessments by describing how reliability is reported for one operational, large-scale diagnostic assessment system: Dynamic Learning Maps (DLM) Alternate Assessments. The DLM assessment system reports student results as mastery of individual skills that are assessed within each content standard and then aggregates the results for broader content strands, providing an overall achievement level for the subject. Correspondingly, reliability evidence is provided for each level of reporting. The methods summarized here can be applied to other diagnostic assessments that require multiple levels of reporting based on stakeholder needs.

Background

Diagnostic Assessments

Diagnostic classification models (DCMs; e.g., Rupp et al., 2010; also known as cognitive diagnosis models) are the basis of diagnostic assessment systems in which students are classified as masters or non-masters of each skill the assessment measures. DCMs may be preferable to traditional psychometric methods of scoring and reporting when fine-grained information about student performance that goes beyond a single raw- or scale-score value is desired.

The foundation of diagnostic assessment systems that use DCMs for scoring and reporting is the specific skills or attributes measuring the construct of interest. Items are written to assess those skills, and the relationship between items and the skills they measure is defined by what is known as the Q-matrix (Tatsuoka, 1983). The Q-matrix is an *n*-item by *m*-skill matrix filled with ones and zeros. A one indicates that an item measures the corresponding skill, and a zero indicates that the skill is not measured by the item. Using this Q-matrix, the DCM estimates posterior probabilities of student mastery for each skill that is assessed. These probability values can be reported as the scores for a diagnostic assessment; however, the desired reporting is usually a dichotomous decision of mastery. Thus, a cut point on the posterior probabilities may be defined, above which students are labeled masters and below which students are labeled non-masters. After applying a cut point to differentiate the two groups, assessment results can be reported as mastery classifications for each skill in the Q-matrix, which allows parents, teachers, and students to use the results to inform next steps for instruction. Skill mastery can also be aggregated to larger grain sizes of reporting by grouping skills together, such as by content strand or overall subject, as is often needed by state accountability programs.

Evaluating Reliability

Reliability indices are calculated to summarize the degree of precision expressed in reported assessment results; they indicate how likely it is across multiple administrations that assessment results will vary because of chance. When reliability indices are high, results are expected to be very consistent from one measurement to the next. In the purest sense, the desired metric is a test–retest correlation without the effects of practice or fatigue; however, this ideal scenario is implausible in practical administrations. It is especially unlikely in large-scale operational assessment systems used for state accountability purposes, when it is impractical to

administer the same assessment (or a parallel form) twice because of policy concerns and the possibility that students may have forgotten or acquired additional knowledge during the time between administrations. Thus, operational assessment programs often summarize reliability evidence by approximating test–retest reliability through other means.

Traditional reliability metrics. Traditional approaches typically quantify reliability of observed scores as a combination of a student's true score and some degree of measurement error. Historically, one of the most widely reported reliability indices is the Guttman–Cronbach alpha (Cronbach, 1951; Guttman, 1945), which estimates internal consistency by quantifying the proportion of true-score variance to observed-score variance. When a test is perfectly reliable (i.e., $\alpha = 1.0$), test-score variation is a solely the results of individual differences in the trait measured by the sampled test takers. When a test is perfectly unreliable (i.e., $\alpha = 0.0$), this variation is purely due to random error.

For assessments that are calibrated and scored using IRT, standard error is directly associated with the latent variable estimate, often referred to as theta. Therefore, assessments can be built to have greater precision at certain points along a continuum (e.g., around cut scores). Technical documentation often reports reliability evidence via the test information function. This conditional standard error of measurement, unlike the Guttman–Cronbach alpha, is a measure of internal consistency that shows the precision at a specific score point, rather than the overall scale.

Notably, these methods assume a continuous latent trait that gives rise to the item responses. However, for diagnostic assessments, the latent trait is categorical: students are either masters or non-masters of each skill. Thus, these traditional methods for reporting reliability are not consistent with the scoring method or the level of reporting of diagnostic systems. Whereas

assessments using CTT and IRT are primarily unidimensional, diagnostic assessments are inherently multidimensional. It is necessary, therefore, to provide scores not only for each dimension (analogous to IRT score reporting) but also for aggregated dimensions. Thus, as indicated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014), and as called for by Sinharay and Haberman (2009), other means for evaluating reliability that are consistent with the scoring method and the levels at which results are reported must be explored.

Diagnostic Classification Model Methods

Instead of placing examinees on a scale-score continuum, diagnostic assessments rely on dichotomous mastery decisions for each measured skill. Therefore, instead of quantifying the precision of the total-score or scale-score value, reliability for diagnostic assessments summarizes the precision of skill mastery classifications. Because DCM results are reported as dichotomous mastery statuses for each skill and not as values on a continuous scale, the consistency of those mastery classifications is inherently more likely to be stable than it is for IRT estimates along a scale score (Templin & Bradshaw, 2013). In other words, instead of replicating an exact estimate among many possible estimates, mastery classification must only distinguish between the two classes to replicate the same result.

DCMs are also unique in that the quantification of error in mastery classifications of discrete skills is inherent in the mastery determination because the results provided by DCMs are the probabilities (p) of mastery estimated for each assessed skill. Because the standard error is equal to $\sqrt{p(1-p)}$, probability values near .5 represent the point of maximum uncertainty (i.e., maximum error) that a student has mastered or not mastered a skill, as shown by $\sqrt{0.5(1-0.5)} = 0.5$. As probability values approach 1 or 0, precision of measurement

increases (i.e., minimum error), representing maximum certainty that a student has mastered or has not mastered a skill, respectively, as shown by $\sqrt{1(1-1)} = 0$ and $\sqrt{0(1-0)} = 0$.

Because results from diagnostic assessments are often reported as the dichotomous mastery status for each skill measured, thresholds are specified to determine the minimum probability to demonstrate mastery. Thresholds that are farther from the point of maximum uncertainty (i.e., .5) reflect greater confidence in the assignment of mastery status (e.g., students are classified as masters of a given skill when their probability of mastery is at or above .8). This process also ensures the stability of mastery classifications as the threshold departs from .5, thus contributing to the overall reliability evidence for the assessment system.

While the probability of mastery (and its standard error) obtained from a DCM provides some certainty about the estimate provided directly by the model, this probability is at the individual student level, rather than the skill level. Classical methods such as Cronbach's alpha do not take into consideration the dichotomous mastery statuses that form the basis of reporting. Instead, overall consistency of the skill can be evaluated by approximating a second administration of the assessment, mimicking test–retest reliability, to compare student performance across multiple replications. In a research application, Templin and Bradshaw (2013) introduced a method for estimating reliability using the posterior probabilities of mastery directly. In this method, the posterior probability of a student mastering a skill should be a constant, assuming that the multiple administrations of the assessment are truly independent. Thus, the probability of mastery across two independent administrations can be calculated as $p \times p$. Similarly, the probability of each mastery/non-mastery profile for a skill can be defined in a 2x2 contingency table, as shown in Table 1.

[Insert Table 1 about here]

In the method proposed by Templin and Bradshaw (2013), this contingency table is calculated for each student and skill, and then the corresponding individual cells are summed across students to provide an aggregated contingency table for each skill. The tetrachoric correlation (Bonett & Price, 2005) derived from the aggregated contingency table serves as the reliability estimate for the skill.

Although this method provides accurate estimates of skill reliability, it does not extend to aggregations of skills, nor is it able to account for the decision reliability that occurs when a cut score is applied to the posterior probabilities of mastery. In practice, reliability often includes an analysis of decision consistency. Decision consistency refers to how reliably respondents are reclassified as masters or non-masters across test administrations.

Roussos et al. (2007) outlined a process to estimate this reliability. Using a Markov chain Monte Carlo procedure, parallel data sets were generated from the calibrated model parameters in the same manner used in posterior predictive model checking. (For a description of posterior predictive model checking methods, see Gelman et al., 2014.) Each parallel data set was then scored, and the skill mastery statuses among the data sets were compared. Thus, it was possible to estimate the rate at which examinees were correctly classified (i.e., mastery status from the simulated data matched mastery status from the observed data), as well as the rate at which two simulated, parallel tests provided the same mastery decision (i.e., test–retest estimate). This approach has several benefits. First, because each of the parallel data sets is scored using the operational calibration, cut points can be applied and included in the reliability calculation. Also, the skill-level mastery scores can be aggregated into composite scores for each data set, allowing the reliability of the composite scores to also be evaluated.

Although one of the key benefits of diagnostic assessment systems is that they report student performance (and the associated precision of measurement) at the level of individual skill mastery instead of as raw-score or scale-score values for a broad construct, many state education agencies rely upon achievement levels for reporting and accountability determinations. Additionally, teachers find aggregated information on broader content strands beneficial when describing student results to parents and use the more fine-grained reporting of skill mastery to plan subsequent instruction (Karvonen, Clark, & Kingston, 2016; Karvonen, Swinburne Romine, Clark, Brussow, & Kingston, 2017). Therefore, diagnostic assessment systems, such as the DLM system, are most helpful when student results are provided at a range of reporting levels. This also means that reliability evidence must be provided for the same range of reporting levels to support interpretation. The sections that follow describe the approach used by one operational large-scale diagnostic assessment system for evaluating reliability at multiple levels that are consistent with the levels used for reporting results.

Dynamic Learning Maps Alternate Assessment System

The DLM Alternate Assessment System is the first large-scale application of a diagnostic assessment system used for statewide accountability purposes. DLM assessments are administered in 18 states to students with the most significant cognitive disabilities who cannot meaningfully access the general education assessment, even with accommodations. Eligibility to take alternate assessments is determined by IEP teams rather than by disability labels. The DLM assessment system offers tests in English language arts (ELA), mathematics, and science in end-of-year-only and through-course assessment models. For exemplary purposes, the discussion in this paper is limited to the ELA assessment for states participating in the through-course model,

which features instructionally embedded assessments during the year, as well as an end-of-year spring assessment.

The basis of the DLM system is an interconnected learning map model that features *nodes* and the connections between them. Each node measures a discrete skill, and the connections between skills indicate the unidirectional ordering of skill acquisition. Nodes in the DLM maps are measured by alternate content standards, which are of reduced breadth and complexity compared to grade-level college- and career-ready standards. Further, to provide all students access to grade-level academic content, each alternate content standard is associated with five skills that represent the alternate content standard at varying levels of depth, breadth, and complexity. For each content standard, there are three precursor skills that lead to the grade-level target and one successor skill for students going beyond alternate grade-level expectations. The availability of multiple skill levels ensures all students are provided access to grade-level content in a way that is most appropriate for the individual student.

A variant of evidence-centered design (Bechard et al., in press) was used to develop items and tests to specifications that would support intended inferences about what students know and can do. Items are categorized in the Q-matrix based on the specifications used during item development such that each item measures a specific skill level for one alternate content standard. Assessments are available to cover the full test blueprint for each grade and subject. Prior to operational administration, items are internally and externally reviewed for bias and sensitivity, content, and accessibility and field tested to evaluate quality.

Because a diagnostic model is used to score the assessment, assessment results produce mastery classifications for every skill that was assessed in each alternate content standard. Latent class analysis (MacReady & Dayton, 1977) is used to obtain the probability that a student

mastered each skill. Latent class models are used due to the large number of skills that are assessed. The number of parameters needed for a simultaneously estimated model introduce operational limitations. The first step in the assessment standard-setting process resulted in a cut-point decision of .8 to classify a student as a master of any skill measured by the assessment. The cut-point decision was based on input from the DLM technical advisory committee and DLM Consortium state partners and was informed by impact data from cuts ranging from .5 to .9. The .8 value was selected because it is slightly greater than one standard deviation above .5 (i.e., the point of maximum uncertainty of mastery status), allowing for reasonable certainty in classifications and for variability in students' responses (Karvonen, Clark, & Nash, 2015). Based on the determined cut-point of .8, students with an estimated posterior probability of mastery less than .8 are not considered masters of a skill, while students with a posterior probability of mastery greater than or equal to .8 are considered masters of a skill.

Individual student score reports for DLM assessments summarize student performance on the assessment at multiple grain sizes, comprising two component parts. The Learning Profile portion of the report summarizes mastery of the specific skills measured for each alternate content standard. The Performance Profile portion of the report provides both a broad summary of the student's results in the subject, including performance within larger content strands that organize the alternate content standards into critical learning domains, and an overall achievement level. As described in the next section, reliability evidence is produced from simulated retest data in which model-specific data are generated for students, based on observed skill mastery and reported at five levels, consistent with DLM score reporting.

Methods for Simulating Retest Data

Given the statistical nature of DCMs and stakeholder desire for multiple levels of reporting, the use of simulation for obtaining a hypothetical second administration and estimating test-retest reliability is a viable alternative to more traditional methods. Following Johnson & Sinharay (2018), we define parallel forms of a diagnostic assessment to be "two tests with the same Q-matrix and identical item parameters" (p. 639). As shown in Roussos et al., (2007), by treating the results from the real-data calibration as true, and using the calibrated model parameters to simulate a parallel second administration, classification consistency indices can be calculated across the true and estimated results and summarized at each of the levels of reporting. The specific steps are as follows:

1. Calibrate the model to get parameters needed for simulating a second administration. Evaluate the model fit of the calibration to ensure valid inferences are made.
2. Draw with replacement a student record from the operational dataset. The student's mastery statuses for each measured skill serve as the true values for the simulation.
3. For each item the student was administered, simulate a new response based on the model-calibrated parameters, conditional on mastery status for the skill.
4. Score simulated responses using the operational scoring procedure,[1] imposing the mastery threshold to determine mastery status. If any additional scoring rules are imposed for operational scoring, apply those here.

[1]While latent class analysis for each skill was used in the present study, any DCM can be used for the scoring procedure. Model selection should be based on information about the assessment design and model-fit analyses.

5. Calculate aggregated composites of skills in accordance with the levels of reporting.

6. Repeat steps 2-5 for 2,000,000 simulated students.

The simulated skill and aggregated composite scores are then compared with the estimated values from the observed data. The degree of agreement between the observed and replicated scores provides a measure of test–retest reliability.

Reporting Reliability

Consistent with the levels at which scores are reported, reliability for DLM assessments is reported at five levels: (a) the classification accuracy of each individual skill, (b) the number of skills mastered within each alternate content standard, (c) the number of skills mastered within each content strand, (d) the number of skills mastered within the subject, and (e) the performance-level classification, which is determined by the number of skills mastered for each subject. The nested aggregation structure of the assessment is visualized in Figure 1.

[Insert Figure 1 about here]

For each level of reliability evidence, measures of association are provided to quantify the precision of measurement. Correlation estimates mirror estimates of reliability from contemporary measures such as the Guttman–Cronbach alpha, resulting in values that are reported on the same scale and that are easy to interpret.

Example Presentation of Results

Results for the DLM assessment obtained from the simulation-based retest method are provided at each level of reporting. As previously stated, the focus of this paper is limited to the ELA assessment for the through-course assessment model. For the complete set of results and more information on the methods specific to the DLM assessment, see Dynamic Learning Maps

Consortium (2017). Because the example results presented here are specific to one testing program, applications of the method and the corresponding summary of results may differ according to the needs of the stakeholders, the design and theory of action of the assessment system, and the levels of results summarized in score reports. Future applications should consider these factors when summarizing reliability results.

The example results are summarized from the finest grain size (i.e., the attribute-level skill) to the largest grain size of reporting (i.e., performance level). Skill-level results are derived from the 2x2 contingency tables of estimated and the observed mastery status for each of the 740 skills measured in the ELA assessment across grades 3–12. Results are reported as the tetrachoric correlation (Bonett & Price, 2005) between true and estimated mastery status, the correct classification rate, and Cohen's kappa (Cohen, 1960). Results can be presented in tabular format, similar to the results provided in Table 2, or in graphical format, as demonstrated in Figure 2. In addition to technical documentation containing the summary information for the 740 skills, the specific reliability evidence for each skill is also made available on the assessment's website, consistent with Standard 2.3 (AERA et al., 2014). Based on Landis and Koch's (1977) recommendation that values of .6 and above are acceptable, the same reporting structure was used for reliability results, whereby values below .6 were combined. In total, 98.3% of results were at or above .6.

[Insert Table 2 about here]

[Insert Figure 2 about here]

Similarly, skill-level results can be grouped together to compare the distribution of indices across relevant categories. For example, traditional assessment programs often report conditional standard errors of measurement to indicate how the precision of measurement differs

along the score continuum. Because diagnostic assessment systems do not report total scores, conditional evidence must be provided in another way. For the DLM assessment in ELA, skills are measured at five levels of different complexity; these levels were established to ensure grade-level content spans the continuum of the skills and abilities seen in the alternate assessment population. Results for each skill level are reported in Figure 3, applying the same three metrics used for the overall skill reliability evidence, including the tetrachoric correlation between true and estimated mastery status, the correct classification rate, and Cohen's kappa.

[Insert Figure 3 about here]

The first aggregation of individual skills is the content standard. For the DLM assessment in ELA, each content standard contains five skills, with one skill at each level of complexity. DLM score reports show the number of skills mastered in each content standard. Therefore, reliability evidence was also provided for the academic content standards themselves. Evidence is summarized as the polychoric correlation between the true and estimated number of skills mastered within each content standard, the correct classification rate, and Cohen's quadratically weighted kappa (Cohen, 1968). A polychoric correlation is used because of the content-standard reliability evidence summarizing the total number of skills the student mastered for that standard, with an implied polytomous ordering among the skills. For DLM assessments in ELA, there are 148 content standards across grades 3–12. A summary of results is provided in both tabular (Table 3) and graphical (Figure 2) forms. Results are reported in full on the DLM website, consistent with Standard 2.3 (AERA et al., 2014).

[Insert Table 3 about here]

At the next increase in grain size, content standards are grouped into content strands as identified on the assessment blueprint. Within each subject, the content standards on which

students are assessed are organized into broad content strands associated with regions of the underlying map structure. Similar to content-standard reliability, content-strand reliability summarizes the agreement of the total number of skills mastered within each content strand. Following the same pattern, the number of skills mastered within each content strand can in turn be aggregated to give the total number of skills mastered for the overall subject. Both content-strand and subject-level results can be reported in a table like Table 4, which summarizes subject reliability by grade. Content-strand reliability can be summarized similarly, with a row for each content strand per grade.

[Insert Table 4 about here]

For content-strand and subject reliability, results are reported as the Pearson correlation between the true and estimated number of skills mastered, the average correct classification rate for the skills mastered, and the average Cohen's kappa for the skills mastered. When compared with the results provided by CTT approaches to reporting reliability, the subject-reliability results are similar to evidence of reliability for total scores[2]. The results summarized in Table 4 indicate that the association between true and estimated number of skills mastered is strong, with values ranging from .919 to .985 across all three metrics. These values indicate that, although there may be minor variations in the total number of skills mastered, a high degree of association between the two values was generally observed between the second simulated administration and the first observed administration. In other words, the results from data generated from the calibrated item parameters (i.e., the simulated parallel test administration) were consistent with the scores calculated from the observed data.

---

[2]Because of the different grain sizes and skill acquisitions needed to master various levels, it is not assumed that skills are on an interval scale of measurement.

At the largest grain size that is reported, DLM score reports describe student achievement in each subject relative to the four performance levels. Table 5 displays the polychoric correlation, correct classification rate, and Cohen's kappa for performance-level reliability for the four achievement levels of DLM assessments. Results across the three metrics showed strong associations, with values ranging from .820 to .983, indicating that, when examining the relationship between the true and estimated responses, students were very likely to be classified into the same overall performance level for the subject.

[Insert Table 5 about here]

Across all reporting levels, reliability results from the DLM assessment indicate a strong association between simulated and true results and support the use of the simulation-based method to provide evidence of reliability at each level at which results are reported. Using this simulation method to assess the reliability of aggregated scores produces results that are initially presented at the finest grain size of reporting (i.e., skill mastery, including different skill types) and then are presented more broadly by aggregating the skills into larger, more meaningful units. Each of these units uses the same procedure and similar metrics of evaluation (e.g., correlation, correct classification rate, and kappa), and each provides useful information for the interpretation of reporting metrics.

Discussion

As the operational use of diagnostic assessment systems becomes more prevalent and extends beyond research applications (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014) to implementation in large-scale assessment systems used for statewide accountability purposes, reliability must be reported in ways that are consistent with the assessment design and the grain size of reported student results. This paper expands upon previous research on simulation-based

reliability methods for diagnostic assessments (e.g., Roussos et al., 2007; Templin & Bradshaw, 2013) by describing how reliability can be reported for diagnostic assessment systems that report results at multiple grain sizes of aggregation. The results presented here further contribute to the operational utility of diagnostic assessments because these methods allow for the summarizing of reliability results for all grain sizes of reporting provided, consistent with the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

As these results demonstrate, reported values for reliability obtained using the simulation-based method may appear higher than the values provided by traditional methods that report reliability for continuous raw- or scale-score values (e.g., when using CTT or IRT); this outcome is consistent with the findings of Templin and Bradshaw (2013). The higher reported values for reliability are the effect of an inherent characteristic of diagnostic assessments: results are based on categorical (and, in this case, dichotomous) latent traits. With only two possible outcomes, results are inherently less likely to fluctuate as much as values along a scale-score continuum do. Furthermore, because reliability is a measure of consistency, an additional strength of diagnostic assessment systems is their ability to produce more-consistent results because of the a priori specification of the mastery threshold. For DLM assessments, the mastery threshold was set at 0.8. Specifying a mastery threshold farther from the point of maximum uncertainty of 0.5 ensures greater certainty in the mastery classification, and a smaller standard error makes classifications less likely to fluctuate across multiple administrations (whether true or simulated). One way that programs that administer diagnostic assessments can ensure greater precision of measurement, both for individual mastery status as well as in higher-level aggregated results, is to specify a higher mastery threshold. Because the specification of the mastery threshold has important implications for reliability, as well as for the larger validity of inferences that can be

made from results, the threshold value should be carefully determined and should be informed by several factors, including feedback from relevant stakeholders.

Results obtained from the simulation-based reliability method are affected by the typical limitations of simulation studies. Because model parameters are used to sample the second set of item responses, evidence of model data-fit is needed to support the use of this method (and the model overall). Item misfit should also be examined and its impact considered prior to reporting reliability because the method uses simulated item responses. Items that do not fit well with the model may affect reported reliability results. Additionally, the use of a second set of simulated responses to approximate test–retest reliability is more computationally intensive than are more traditional methods for calculating reliability. Operational assessment programs using a simulation-based method will need to factor these additional time constraints into the timeline for reporting reliability evidence of assessment results.

As an additional caution, the reliability estimates reported here represent an upper bound of the true reliability and are contingent on the fit of the model to the data. Thus, test design and subsequent model fit analyses are critical to the final reported reliability estimates (for a summary of model fit for the DLM assessment see Dynamic Learning Maps Consortium, 2017, Chapter V). Assessments scored with a DCM should be designed for diagnostic purposes to ensure accurate interpretation of reliability estimates. Retrofitting diagnostic models to assessments designed to measure a single continuous latent trait can have important implications for the interpretation of results and of corresponding reliability estimates, and is generally not advisable.

While the methods described here provide an approach for summarizing reliability evidence at each level of reporting for technical documentation purposes, additional research is

needed into how best to report precision of measurement on the score reports themselves, in accordance with the *Standards for Educational and Psychological Testing* (AERA et al., 2014). For example, the probability of mastery of individual skills could be indicated via shading using a color spectrum (e.g., red to green continuum representing certainty of mastery; see Rupp et al., 2010, for an illustrated example). As the use of diagnostic assessment systems expands, additional research into parent and teacher interpretation of reported mastery should be conducted to determine the best ways to present this information.

References

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

Bechard, S., Clark, A. K., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (in press). Using evidence-centered design to develop learning maps-based assessments. *International Journal of Testing*.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30,* 213–225. https://doi.org/10.3102/10769986030002213

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2–14. doi:10.1111/emip.12020

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. http://doi.org/10.1111/emip.12162

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

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

Dynamic Learning Maps Consortium. (2017). *2016–2017 Technical Manual Update: Integrated Model.* Lawrence: University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). Boca Raton, FL: CRC Press.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: SAGE.

Karvonen, M., Clark, A. K., & Kingston, N. (2016, April). Designing alternate assessment score reports: Implications for instructional planning. In P. Kannan (Chair), *Thinking about your audience in designing and evaluating score reports.* Symposium presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Karvonen, M., Clark, A. K., & Nash, B. (2015). *2015 Year-End Model Standard Setting: English Language Arts and Mathematics* (Technical Report No. 15-03). Lawrence: University of Kansas, Center for Educational Testing and Evaluation.

Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., & Kingston, N. (2017, April). *Promoting accurate score report interpretation and use for instructional planning.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. http://www.jstor.org/stable/2529310

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. New York, NY: Cambridge University Press.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Oxford, England: Addison-Wesley.

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120. www.jstor.org/stable/1164802

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, NY: Cambridge University Press.

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

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 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

Table 1

Contingency Table for Mastery and Non-Mastery of a Single Skill Across Two Hypothetical Administrations of an Assessment

	Administration 2	
Administration 1	Master	Non-master
Master	$p \times p$	$p(1-p)$
Non-master	$(1-p)p$	$(1-p)(1-p)$

Table 2

Number of Skills in Each Range for the Reported Agreement Statistics

	Index range								
Metric	<.60	.60.–64	.65–.69	.70–.74	.75–.79	.80–.84	.85–.89	.90–.94	.95–1.00
Tetrachoric correlation	2	2	0	1	8	9	35	100	583
Correct classification rate	0	0	0	0	0	7	84	316	333
Cohen's kappa	35	21	17	47	89	173	168	92	98

Table 3

Number of Content Standards in Each Range for the Reported Agreement Statistics

	Index range								
Metric	<.60	.60–.64	.65–.69	.70–.74	.75–.79	.80–.84	.85–.89	.90–.94	.95–1.00
Polychoric correlation	0	0	0	0	1	14	32	81	20
Correct classification rate	0	0	0	4	16	58	57	13	0
Cohen's kappa	0	0	1	3	8	20	59	52	5

Table 4

Subject Reliability, by Grade

Grade	Skills mastered correlation	Average student correct classification	Average student Cohen's kappa
3	.981	.982	.963
4	.983	.984	.966
5	.979	.978	.952
6	.976	.974	.943
7	.964	.965	.919
8	.971	.968	.927
9	.980	.977	.948
10	.980	.977	.947
11	.974	.967	.923
12	.969	.985	.964

Table 5

Performance-Level Reliability, by Grade

Grade	Polychoric correlation	Correct classification	Cohen's kappa
3	.983	.858	.930
4	.979	.892	.939
5	.983	.867	.930
6	.981	.858	.918
7	.970	.838	.893
8	.976	.827	.914
9	.983	.850	.927
10	.983	.851	.930
11	.974	.820	.908
12	.983	.905	.917

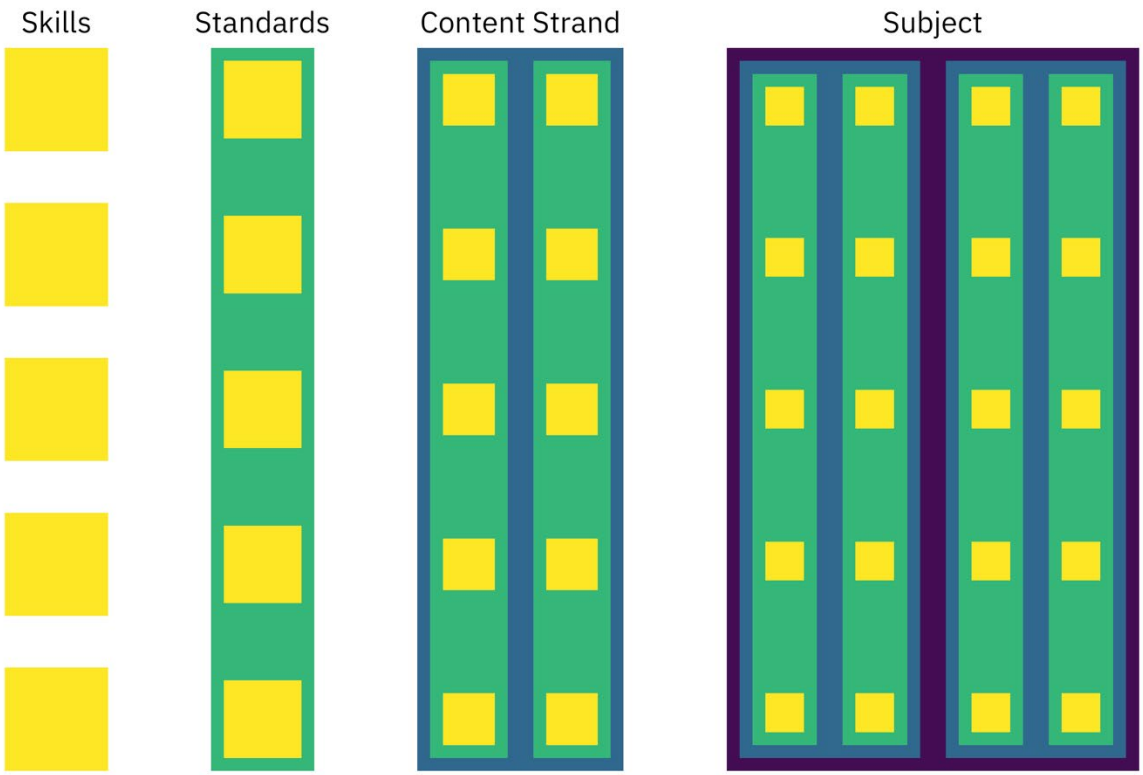

Figure 1. Aggregation structure of the DLM assessment. Skills are nested within alternate content standards, content standards within content strands, and content strands within the overall subject.

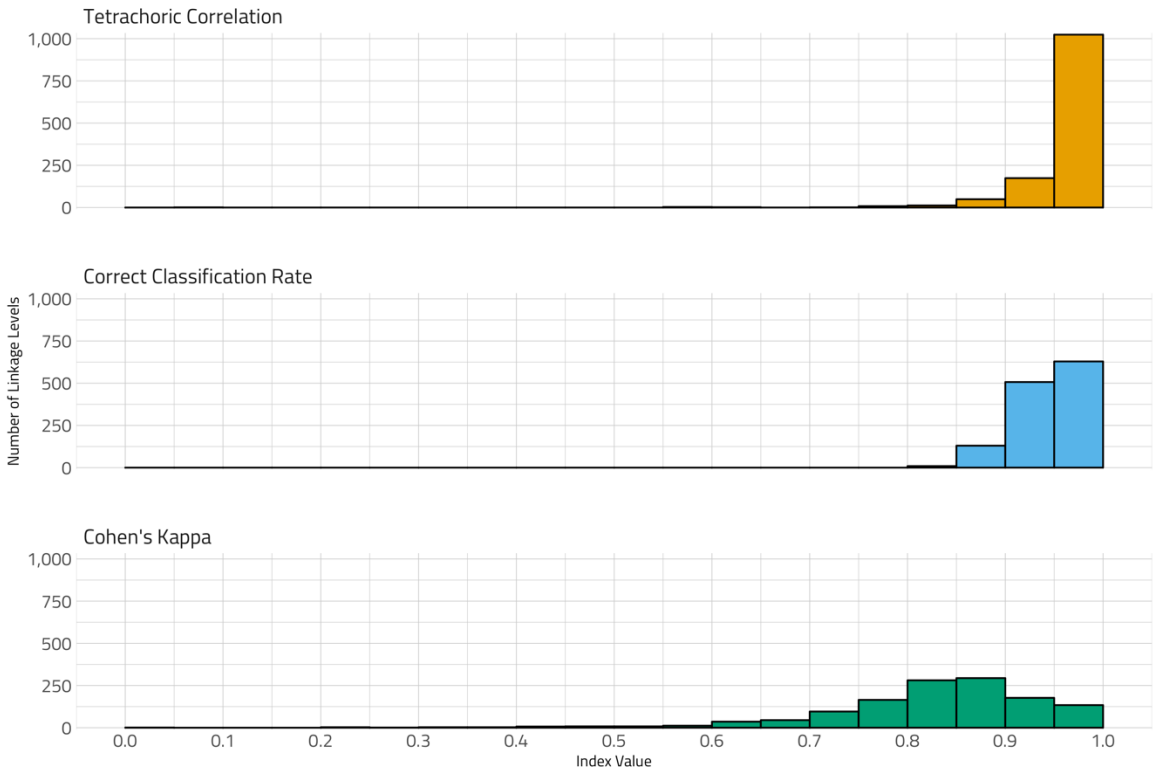

Figure 2. Distributions of agreement statistics for all skills.

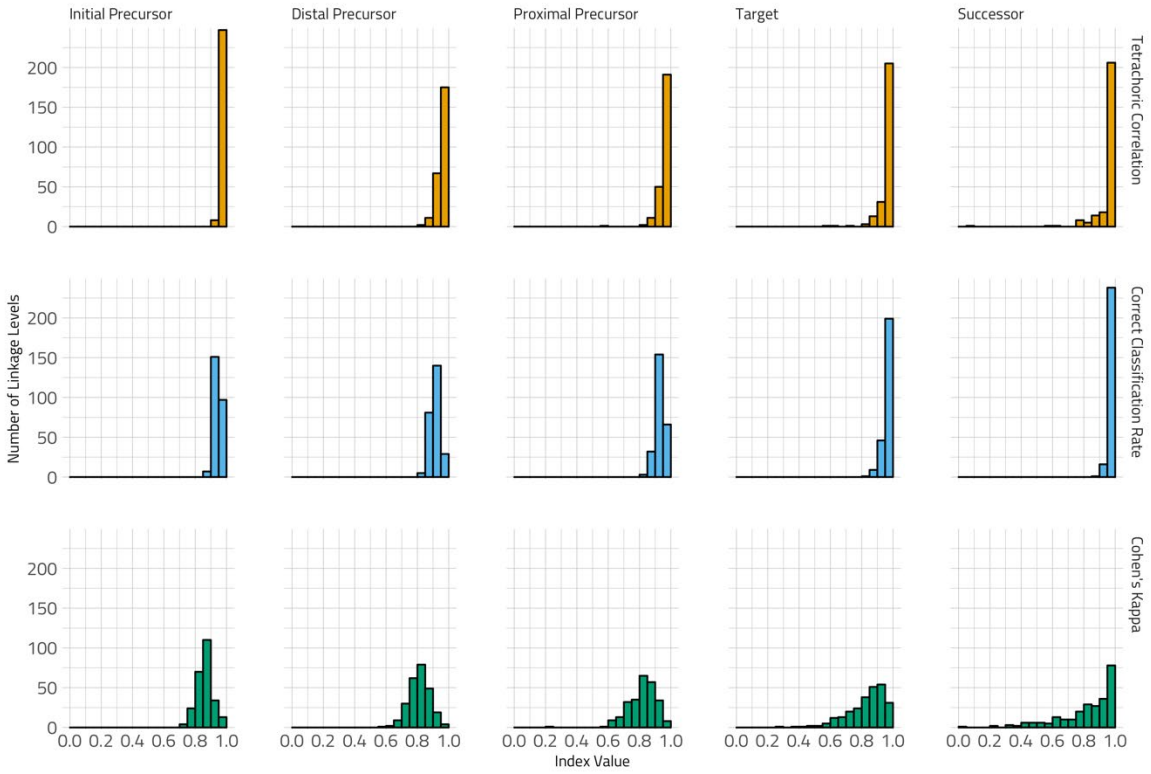

Figure 3. Distributions of agreement statistics for skills, by level of complexity.