DYNAMIC® LEARNING MAPS

Reliability for the Dynamic Learning Maps Assessments: A Comparison of Methods

Technical Report #20-03

July 2020

Thompson, W. J. (2020). *Reliability for the Dynamic Learning Maps Assessments: A comparison of methods* (Technical Report No. 20-03). University of Kansas, Accessible Teaching, Learning, and Assessment Systems.

Contents

List of Tables

List of Figures

Executive Summary

Dynamic Learning Maps® (DLM®) alternate assessments use diagnostic classification models (DCMs) to report the knowledge, skills, and understandings of students with the most significant cognitive disabilities. To meet the needs of various stakeholders, student achievement is reported at multiple levels of aggregation. Reported results include mastery classifications for each individual skill, as well as aggregations of mastered skills within alternate content standards (Essential Elements [EEs]), content strands (conceptual areas/claims/domains), subject areas, and an overall performance level for the subject. This reporting structure ensures that fine-grained information is available to help teachers target specific instructional goals while also providing a high-level overview of student achievement that is often necessary for state accountability systems.

Because results are reported at multiple levels, the reliability of each level of scoring must also be evaluated. To assess reliability, DLM assessments use an innovative simulated retest methodology. This method works by first generating a hypothetical second administration of the DLM assessment for students, and then comparing the results across the observed and simulated test administrations. Although prior work has provided theoretical support for this method, the reliability estimates from the simulation methodology have not been compared empirically to other estimates of reliability used in the DCM literature. This report describes a comparison between the simulation methodology for reliability used for DLM assessments and popular non-simulation approaches to evaluate reliability for DCMs.

The key findings from this report are:

- Reliability estimates from the simulation method are largely consistent with estimates from non-simulation approaches.
- Discrepancies between simulation and non-simulation methods are likely due to low sample sizes and unbalanced mastery classifications.

The findings provide evidence that the simulation methodology used for DLM assessments provides reliability estimates that are consistent with traditional approaches. The non-simulation estimates all fall within the ranges of estimates provided by the summaries of the simulated retests, indicating that the simulation method is systematically unbiased.

1. Purpose of the Report

The Dynamic Learning Maps® (DLM®) alternate assessments report student achievement as a profile of mastery on the set of discrete skills in English language arts (ELA), mathematics, and science. The profiles of skill mastery reported on DLM summative score reports are able to provide fine-grained information to teachers and parents about specific areas where their students are performing well and where extra attention may be needed. This is in contrast to more conventional assessments, which use a single scale score for reporting assessment outcomes. To provide the mastery profiles, DLM assessments are scored using a diagnostic classification model (DCM; Bradshaw, 2016; Rupp et al., 2010). Although DCMs have been widely used in the research literature, their use in applied settings has been limited (Ravand & Baghaei, 2020; Sessoms & Henson, 2018). As a result, the use of DCMs in operational assessment settings requires innovation in order to provide the necessary technical evidence to support the intended uses of the assessment. One piece of this technical evidence is reliability.

The *Standards for Educational and Psychological Testing* defines reliability “in terms of consistency over replications of the testing procedure” (American Educational Research Association et al. [AERA et al.], 2014, p. 35) and specifies that reliability evidence should support “interpretation for each intended score use” (AERA et al., 2014, p. 42). For DLM assessments, this means not only reporting the reliability of the individual skills, but also aggregated summaries of skills used for state accountability reporting (e.g., skills within conceptual areas, overall performance level achievement). To meet this need, DLM assessments assess reliability by simulating repeat administrations of the assessment and compare the results from the simulated retest to the observed results. The simulation procedure meets the needs for evaluating the consistency over replicated tests and provides flexibility for reporting reliability at multiple reporting levels (see Thompson et al., 2019). However, the reliability estimates from this procedure have not yet been compared to estimates from non-simulation-based reliability methods for DCMs that have been used in the research literature. In this report, we briefly describe the structure of DLM assessments and the simulation method utilized to evaluate their reliability, as well as non-simulation approaches to reliability for DCMs. All methods are then applied to operational DLM data to evaluate how the estimated reliability differs across the methods.

2. The Dynamic Learning Maps Alternate Assessments

DLM alternate assessments measure what students with the most significant cognitive disabilities know and can do relative to grade-level academic expectations in ELA, mathematics, and science. For each subject, the test blueprint specifies the alternate content standards (called Essential Elements [EEs]) for each grade, which are organized into overarching conceptual areas and claims or domains. To provide all students access to grade-level academic content, each EE is measured at multiple skills called linkage levels. Linkage levels are collections of related nodes in the underlying map structure, which is the conceptual and content basis for the DLM assessments.[1] In ELA and mathematics, there are five linkage levels for each EE: the Target level, which represents the grade-level expectations; three precursor levels (Initial Precursor, Distal Precursor, and Proximal Precursor) that lead up the Target; and the Successor level, which is available for students extending beyond the Target. In science, there are three linkage levels. As in ELA and mathematics, the Target level represents the grade-level expectations; however, in science, there are only two precursor levels (Initial and Precursor) leading up to the Target and no levels extending beyond the Target. The assessment is delivered in the form of testlets, which generally measure a single EE and linkage level and consist of 3–5 items.[2]

For ELA and mathematics, there are two assessment models available to state partners that determine how testlets are delivered. These models are the Instructionally Embedded and the Year-End assessment models. The DLM science assessment follows the same delivery structure as the Year-End assessment for ELA and

[1] For a full description of the learning map model, see Chapter 2 of the *2014–2015 Technical Manual–Integrated Model* (Dynamic Learning Maps Consortium [DLM Consortium], 2016a).

[2] On ELA assessments, all writing EEs are delivered on a single testlet with up to nine items.

mathematics. The Instructionally Embedded assessment model features two testing windows, one in the fall and the other in the spring. In both windows, test administrators choose which EEs to assess (within constraints to ensure adequate breadth of blueprint coverage), as well as which linkage level each EE is assessed at. Students are expected to cover the full test blueprint in both the fall and spring assessment windows, and data from both windows are used for summative results. In 2018–2019, there were five states that participated in the Instructionally Embedded assessment model for ELA and mathematics.

The Year-End assessment model includes only one required testing window in the spring. The fall testing window is available to Year-End model states, but participation is optional, and data from the fall window are not included in summative results. During the Year-End spring windows, all EEs on the test blueprint are assessed, and the linkage level for each EE is determined by the system. The linkage level for the first testlet is determined by responses to the First Contact survey, which is a survey of learner characteristics completed by the student's teacher. Subsequent testlets are determined by an adaptive routing algorithm. Specifically, the linkage level associated with the next testlet received by a student is based on the student's performance on the previous testlet. The system adapts up one linkage level if the student answers at least 80% of items correctly on the previous testlet. The system adapts down one linkage level if the student answers less than 35% of items correctly on the previous testlet. The system stays at the same linkage level if the student answers between 35% and 80% of items correctly. This process of a student completing a testlet and the system adapting to determine the linkage level of the next testlet continues until the full test blueprint has been covered. For a full description of the adaptive algorithm, see Chapter 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016b). In 2018–2019, there were 13 states that participated in the Year-End assessment model for ELA and mathematics. Additionally, there were 17 states that participated in the DLM science assessment, which follows the same process for testlet assignment as the Year-End model for ELA and mathematics.

Once the assessment is administered, the linkage level is used as the unit of scoring for DLM assessments. That is, a mastery probability is estimated for each linkage level the student was assessed on. In DCM terminology, the linkage levels are the attributes in the estimated models. Students are then assigned a mastery status for each linkage level based on the mastery probability and other scoring rules. Specifically, students are classified as a master of the linkage level if their posterior probability of mastery for the linkage level is greater than .8 or if the student answered at least 80% of the items measuring the linkage level correctly. Additionally, mastery is assumed for all linkage levels below the highest mastered linkage level. That is, if a student demonstrates mastery of the Target linkage level for a mathematics EE, then they are also assigned mastery of all preceding linkage levels (i.e., Initial Precursor, Distal Precursor, and Proximal Precursor). If a student does not meet either threshold on any linkage level assessed for an EE, mastery is assigned at two linkage levels below the lowest tested linkage level. Using the same example of a mathematics Target linkage level, if the student did not meet the posterior probability or percentage correct threshold, they would be assigned mastery of the Distal Precursor (i.e., two levels below the tested linkage level). For details on the psychometric model and scoring procedures, see Chapter 5 of the *2015–2016 Technical Manual Update—Integrated Model* (DLM Consortium, 2017).

The individual linkage level mastery statuses are then aggregated into the number of levels mastered within an EE; conceptual area, claim, or domain; and an overall performance level for the subject. The linkage level mastery and aggregated mastery statuses are both included on summative score reports.

3. Simulated Reliability for Dynamic Learning Maps Assessments

The simulation procedure used for estimating the reliability of DLM assessments is described in detail in Chapter 8 of the *2015–2016 Technical Manual Update—Integrated Model* (DLM Consortium, 2017) and Thompson et al. (2019). A high-level overview of the procedure is provided here.

3.1. Simulation Method for Reliability

The simulation procedure is designed to estimate test-retest reliability by simulating a hypothetical second test administration. As shown in Roussos et al. (2007), calibrated model parameters can be used to simulate a parallel test administration. Following Johnson and Sinharay (2018), parallel forms for a diagnostic assessment are defined as “two tests with the same Q-matrix and identical item parameters” (p. 639). Classification indices can then be calculated across the observed and estimated results to summarize results at each level of reporting (Thompson et al., 2019). The specific steps for the simulation procedure implemented for DLM assessments are as follows:

1. Draw with replacement a student record from the operational assessment data.
2. Assign the mastery status for each linkage level randomly, with the probability of being assigned to the master class equal to the observed mastery probability from the observed data.
3. For each testlet taken by the student, simulate new responses for that EE using calibrated model parameters. The linkage level of the simulated testlets may be different than the observed linkage level if adaptation decisions between simulated testlets were different than the decision between observed testlets.[3]
4. Score simulated responses using the operational scoring procedure, imposing the mastery threshold and any additional scoring rules to determine mastery status.[4]
5. Calculate aggregated composites of linkage levels consistent with the levels of score reporting.
6. Repeat steps 1–5 for 100,000 simulated students in each grade and subject.

The estimated mastery classifications and aggregated results from the simulated retest administrations are then compared to the corresponding values derived from the observed data. The degree of agreement between the observed and simulated values represents a measure of test-retest reliability.

3.2. Reporting Reliability for Dynamic Learning Maps

There are many measures that can be used to define the degree of agreement between the observed and simulated values. The decision for the reported summary statistics should be based on the levels of reported results as well as the design and theory of action of the assessment system. Because non-simulation-based approaches are unable to estimate reliability at aggregated levels of reporting, this report will focus on the metrics used for reporting the reliability of the linkage level (attribute) mastery classifications.

For DLM linkage levels, three metrics are reported to summarize the agreement between the observed and simulated mastery classifications.[5] The first is percent classification agreement. The percent classification agreement is simply the proportion of linkage level mastery classifications that are the same across the observed administration and simulated retests. This method provides the clearest evaluation of agreement between observed and simulated mastery classifications; however, this measure may be biased for linkage levels where there is a high (or low) rate of mastery. For example, if 90% of students are masters of a given linkage level, we would expect a high rate of agreement just from chance, even if there was no relationship between observed and simulated classifications. The second and third metrics used to summarize the agreement between the observed and simulated mastery classifications account for chance agreement to address this limitation of the percent classification agreement.

The second metric used to summarize linkage level reliability for DLM assessments is the tetrachoric correlation

[3] For a description of adaptive delivery for DLM assessments, see Chapter 4 of the *2014–2015 Technical Manual—Year-End Model* (DLM Consortium, 2016b).

[4] For a description of the DLM scoring rules, see Chapter 5 of the *2015–2016 Technical Manual Update—Integrated Model* (DLM Consortium, 2017).

[5] For a summary of the DLM reliability evidence for linkage levels and other levels of reporting, see Chapter 8 of the *2015–2016 Technical Manual Update—Integrated Model* (DLM Consortium, 2017).

between the observed and simulated mastery classifications. The tetrachoric correlation is the association between two latent random normal variables that results in quadrant probabilities equal to the proportions of a 2 × 2 contingency table of observed and simulated mastery classifications (Kirk, 1973; Pearson, 1900). By approximating the classification contingency table, the tetrachoric correlation is able to account for the base rate of mastery class membership (Banerjee et al., 1999; Warrens, 2008). However, Hripcsak and Heitjan (2002) notes that the tetrachoric correlation is intended for a latent trait (i.e., continuous latent variable) and is not necessarily intended to be used in a latent class model (such as DCMs).

The third metric used to summarize the reliability of linkage levels is Cohen's kappa (Cohen, 1960). Like the tetrachoric correlation, Cohen's kappa assesses whether the observed agreement exceeds what would be expected due to chance; however, unlike the tetrachoric correlation, Cohen's kappa does not rely on the assumption of an underlying continuous latent trait (Agresti, 1992). When there are more than two categories (e.g., for examining aggregated EE reliability), the quadratically weighted kappa is used (Cohen, 1968). The weighted Cohen's kappa is calculated as shown in Equation 1, where k is the number of categories, x is the observed contingency matrix, m is the expected contingency matrix, and w is the matrix of weights. If all off-diagonal elements of w are equal to one, or if there are only two categories ($k = 2$), Equation 1 is equivalent to the unweighted kappa.

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}} \tag{1}$$

Because Cohen's kappa doesn't rely on a continuous underlying latent trait, it may provide a better estimation of linkage level reliability in some cases. However, the kappa statistic also has limitations. Several studies have noted that the kappa statistic is overly conservative when the categories are unbalanced (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; O'Leary et al., 2014; Pontius & Millones, 2011). This can sometimes be the case for DLM linkage levels, where the base rate of linkage level mastery is often observed to be less than .30 or greater than .70.[6] Thus, it is important to interpret the kappa statistic with caution and within the context of the other reporting metrics.

In summary, each metric has strengths and limitations that are balanced by the strengths and limitations of the other metrics. Thus, the set of metrics provides an overall evaluation of the reliability of DLM linkage levels. Although the focus here is on metrics for reporting the reliability of linkage levels, similar metrics are reported for each level of aggregated reporting as well. At each level, three metrics are provided: the percent classification agreement, a measure of correlation (i.e., tetrachoric, polychoric, or Pearson), and Cohen's kappa. For more information on the aggregated levels of reporting, see Thompson et al. (2019).

4. Non-Simulation Approaches to Reliability

For DCM-based assessments that are not used for state accountability systems, results are typically reported as either the probability that each attribute was mastered (expected *a posteriori*) or the dichotomous mastery classification (maximum *a posteriori*). Accordingly, the reported reliability should reflect the type of results that are reported, as the reliability of a mastery probability may be quite different from the reliability of a mastery classification. Assessments using the mastery probability generally report the consistency of the posterior probability of mastery (e.g., Johnson & Sinharay, 2019; Templin & Bradshaw, 2013), whereas assessments using mastery classifications generally report the classification consistency and/or accuracy (e.g., Cui et al., 2012; Johnson & Sinharay, 2018; Wang et al., 2015). Although there are many methods for estimating the reliability of DCM-based assessments at both the mastery profile and attribute levels, only a brief overview of methods applicable to DLM assessments is provided here. Specifically, because the DLM assessments calibrate and

[6] For details on the base rate of linkage level mastery, see Chapter 5 of the *2018–2019 Technical Manual Update—Integrated Model* (DLM Consortium, 2019).

estimate mastery for each attribute (i.e., linkage level) separately, we focus here on the attribute-level indices. For a comprehensive overview of DCM reliability methods, see Sinharay and Johnson (2019). Consistent with the results reported for DLM assessments and the recommendations from the *Standards for Educational and Psychological Testing* for reporting consistency in results, we focus on measures of classification consistency in this report.

At the highest level, classification consistency is the probability that a student would receive the same mastery classification for attribute d on two parallel forms of the assessment. Thus, classification consistency can be defined as shown in Equation 2, where $\tilde{a}_d$ is the estimated mastery classification, and $\mathbf{X}$ is the vector of item responses from an administration of the assessment.

$$\mathrm{P}_{\mathrm{CC}_d} = \mathrm{P}(\tilde{a}_d(\mathbf{X}_1) = \tilde{a}_d(\mathbf{X}_2)) \tag{2}$$

Wang et al. (2015) suggested estimating $\mathrm{P}_{\mathrm{CC}_d}$ using the estimator $\hat{\gamma}_d$, as defined in Equation 3, where A_d is the true mastery classification for attribute d, and N is the total number of students (Wang et al., 2015, Equation 9).

$$\hat{\gamma}_d = \frac{1}{\mathrm{N}} \sum_{n=1}^{\mathrm{N}} \left([\mathrm{P}(\mathrm{A}_d = 1 | \mathbf{X} = \boldsymbol{x}_n)]^2 + [\mathrm{P}(\mathrm{A}_d = 0 | \mathbf{X} = \boldsymbol{x}_n)]^2 \right) \tag{3}$$

In Equation 3, $\hat{\gamma}_d$ approaches 1.0 as the probability of attribute mastery approaches 0 or 1. Intuitively, this follows what would be expected from a measure of classification consistency. That is, as the probability of mastery approaches 0 or 1, the classification becomes more certain, and therefore should be expected to be more consistent across parallel test forms. Implicit in the $\hat{\gamma}_k$ estimator proposed by Wang et al. (2015) is the assumption that the posterior probability of mastery for a given respondent and attribute is constant across test forms that are parallel. However, Johnson and Sinharay (2018) show that this assumption is invalid except in unrealistic and extreme instances where non-masters and masters have a probability of providing a correct response of 0 and 1, respectively.

Johnson and Sinharay (2018) propose an alternative estimator for $\mathrm{P}_{\mathrm{CC}_d}$ based on the 2 × 2 contingency table formed from the mastery classifications across two parallel test administrations, such as in Table 1. In the contingency table, each r_{ij} is the joint probability of mastery classifications, shown in Equation 4.

$$r_{ij} = \mathrm{P}(\tilde{a}_d(\mathbf{X}_1) = i,\ \tilde{a}_d(\mathbf{X}_2) = j) \tag{4}$$

Table 1

Contingency Table for Classification Consistency

	Estimate from Form 2 $(\tilde{a}_d(\mathbf{X}_2))$		
Estimate from Form 1 $(\tilde{a}_d(\mathbf{X}_1))$	0	1	Total
0	r_{00}	r_{01}	r_{0+}
1	r_{10}	r_{11}	r_{1+}
Total	r_{+0}	r_{+1}	1

To estimate $\hat{r}_{ij}$ from a single test administration, Johnson and Sinharay (2018) provide the estimator shown in Equation 5, where Ω is the set of all possible attribute mastery patterns (that is, all values of $\mathbf{A}$) and $\mathbf{A}$ represents a profile of mastery across assessed attributes (Johnson & Sinharay, 2018, Equation 25).

$$\hat{r}_{ij} = \sum_{\boldsymbol{a}\in\Omega} \frac{\left(\sum_{n=1}^{\mathrm{N}} \mathrm{P}(\mathbf{A}=\boldsymbol{a}|\mathbf{X}=\boldsymbol{x}_n)\mathrm{I}\{\hat{a}_{nd}=i\}\right)\left(\sum_{n=1}^{\mathrm{N}} \mathrm{P}(\mathbf{A}=\boldsymbol{a}|\mathbf{X}=\boldsymbol{x}_n)\mathrm{I}\{\hat{a}_{nd}=j\}\right)}{\mathrm{N}^2\mathrm{P}(\mathbf{A}=\boldsymbol{a})} \tag{5}$$

The alternate estimator of $\mathrm{P}_{\mathrm{CC}_d}$ suggested by Johnson and Sinharay (2018), denoted $\hat{\mathrm{P}}_{cd}$, can then be calculated from the estimated contingency table.

$$\hat{\mathrm{P}}_{cd} = \hat{r}_{00} + \hat{r}_{11} \tag{6}$$

Both the method proposed by Wang et al. (2015) and by Johnson and Sinharay (2018) are limited in that they do not generalize to multiple levels of score reporting the way the simulation methodology used for DLM assessments is able to (Thompson et al., 2019). However, it is possible to compare the values of $\hat{\gamma}_k$ and $\hat{\mathrm{P}}_{cd}$ to the estimates of linkage level reliability from the simulation method. In the next section, we apply the methods of Wang et al. (2015) and Johnson and Sinharay (2018) to DLM linkage levels and compare the resulting reliability estimates to those achieved with the simulation method.

5. Comparison of Reliability Methods

Although there is a strong theoretical foundation supporting the simulation methodology for estimating the reliability of individual attributes, this method has not yet been compared to more traditional non-simulation approaches. To investigate the similarities and differences between the simulation and non-simulation approaches, each method was applied to each linkage level included in the DLM assessment. Specifically, the following indices were calculated for each linkage level:

1. Percent classification agreement, from simulated retests
2. Tetrachoric correlation, from simulated retests
3. Cohen's kappa, from simulated retests
4. $\hat{\gamma}_d$, from Wang et al. (2015)
5. $\hat{\mathrm{P}}_{cd}$, from Johnson and Sinharay (2018)

Across all subjects, DLM assessments consist of 1,377 linkage levels, which includes 740 from ELA, 535 from mathematics, and 102 from science. The results from the simulated retests were then directly compared to the corresponding $\hat{\gamma}_d$ and $\hat{\mathrm{P}}_{cd}$ values.

5.1. Data

To compare the attribute-level reliability indices used for DLM assessments to the indices proposed by Wang et al. (2015) and Johnson and Sinharay (2018), each method was applied to the 2018–2019 operational assessment data. For this analysis, the data were limited to states participating in the Instructionally Embedded assessment model for ELA and mathematics.[7] All states participating in the DLM science assessment were included. This resulted in 69,657 assessment administrations being included in the analysis, where one administration is the complete set of testlets for a student in a given grade and subject. Table 2 shows the number of assessment administrations by grade and subject. Variations in high school ELA and mathematics and science participation are due to state-level policies about the grades in which students are assessed.

[7] States participating in the Year-End assessment model adopted blueprint changes beginning with the 2019–2020 assessment. This change would limit the generalizability of results from the Year-End model going forward.

Table 2

Assessment Administrations Included in Reliability Estimation, by Grade and Subject

Grade	English Language Arts	Mathematics	Science
3	1,978	1,974	650
4	1,924	1,920	4,191
5	2,011	2,010	7,391
6	2,023	2,023	748
7	2,015	2,009	718
8	1,873	1,871	10,748
9	959	960	4,385
10	1,447	1,443	1,697
11	1,451	1,445	6,793
12	257	242	297
Biology	—	—	204

The calibrated model parameters used for generating the simulated retests and calculating the Wang et al. (2015) and Johnson and Sinharay (2018) indices came from the 2018–2019 operational scoring model. This calibration was trained on all operational testlets, using data from the 2015–2016 through 2017–2018 assessment administrations. For a summary of the calibrated model, see Chapter 5 of the *2018–2019 Technical Manual Update—Integrated Model* (DLM Consortium, 2019).

5.2. Results

Figure 1 and Table 3 provide the number and proportion of linkage levels, respectively, that fall within pre-specified ranges of values for each of the reliability indices. In Figure 1, the distributions of the Wang et al. (2015) and Johnson and Sinharay (2018) indices are similar and most closely resemble the distributions of the tetrachoric correlation and percent classification agreement summary statistics. The distribution of Cohen's kappa statistics from the simulated retests is shifted slightly lower than the other indices. Table 3 shows the proportion of linkage levels that fall within the pre-specified ranges for each index. Across all linkage levels, 0 had a percent classification agreement less than .6, 22 had a tetrachoric correlation less than .6, 565 had a Cohen's kappa less than .6, 23 had a $\hat{\gamma}_d$ less than .6, and 83 had a $\hat{\mathrm{P}}_{cd}$ less than .6.

Figure 1

Distributions of Reliability Indices across All Linkage Levels

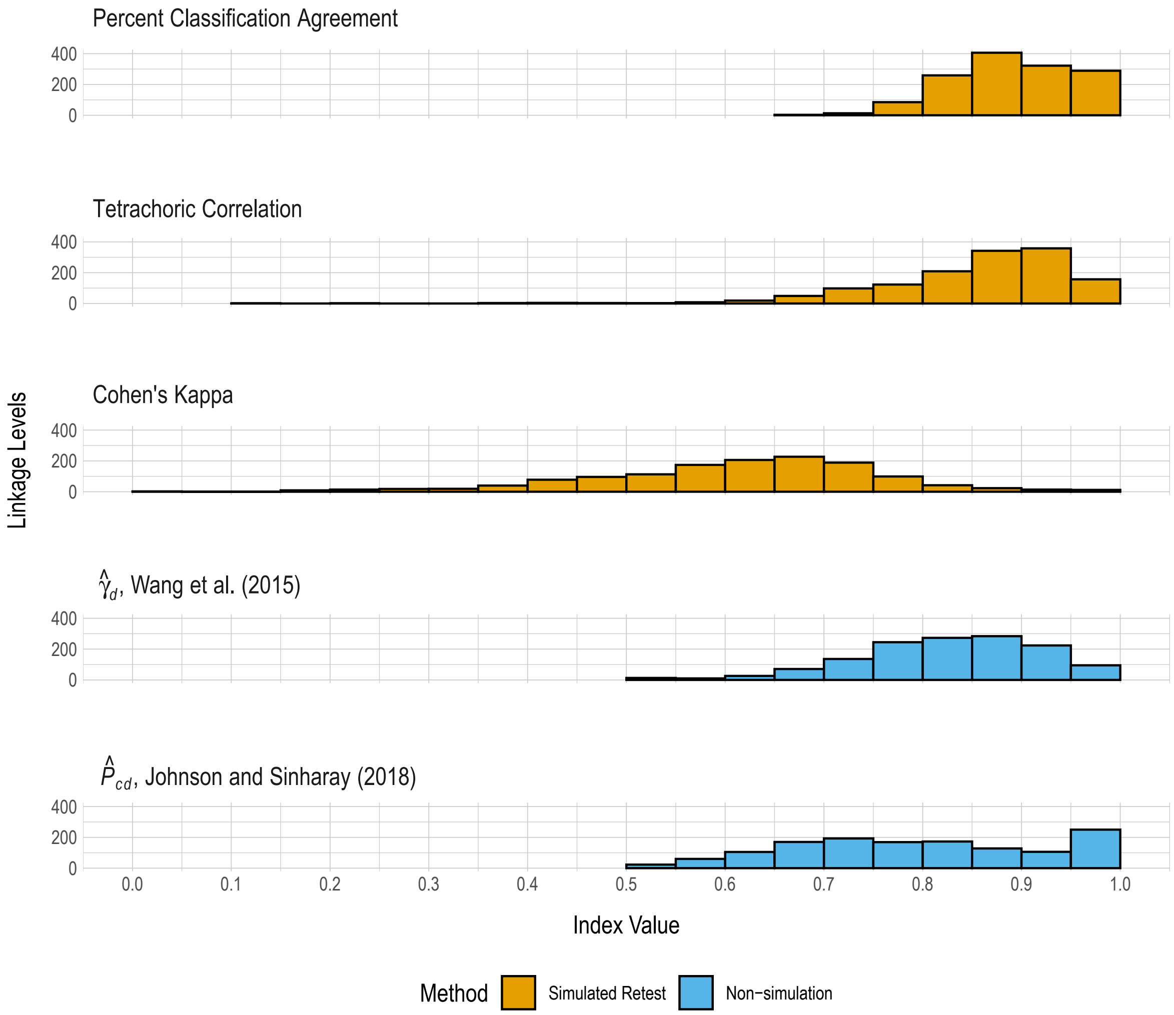


Table 3

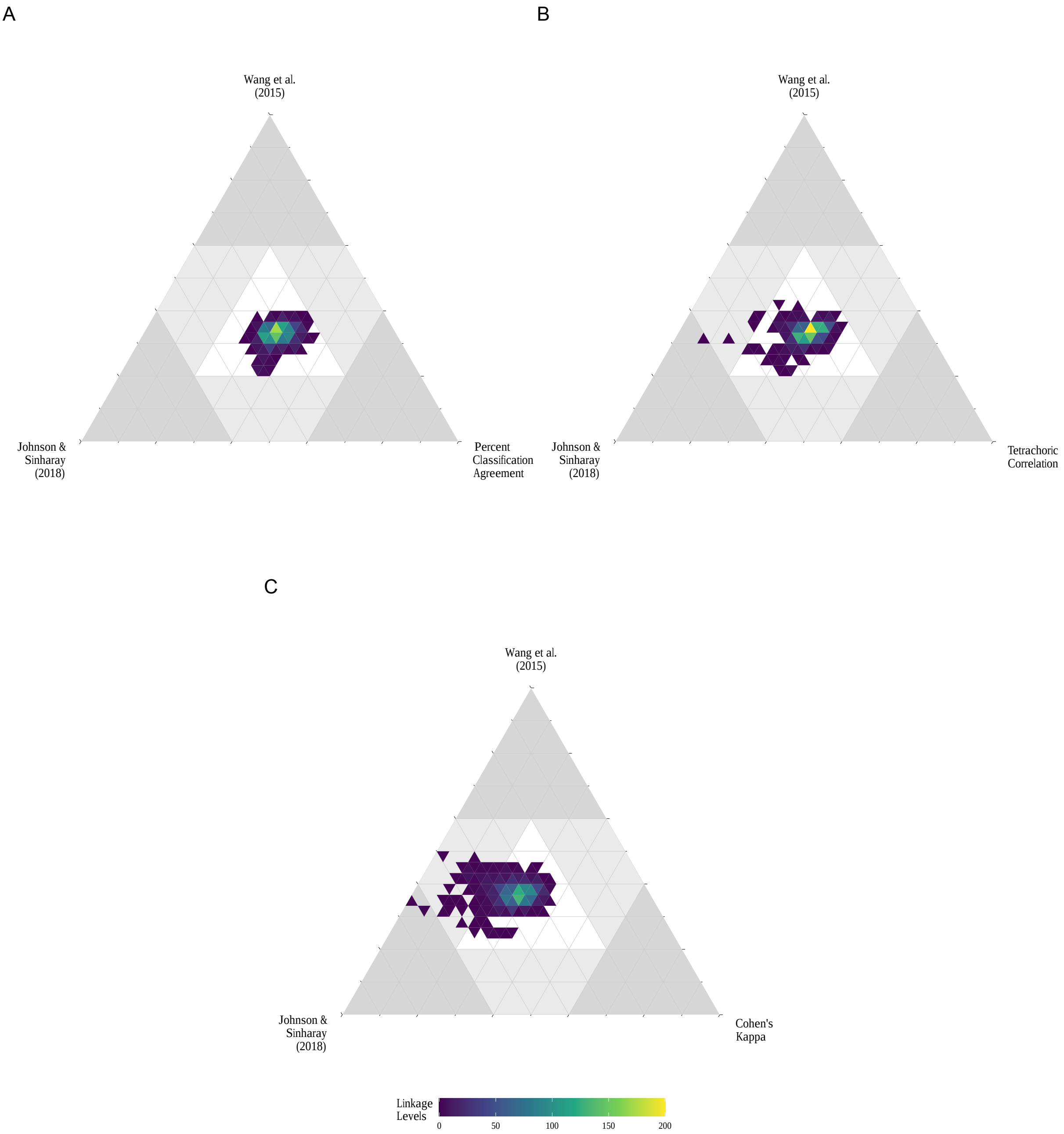
Reliability Summaries Across All Linkage Levels: Proportion of Linkage Levels Falling Within a Specified Index Range

	Index Range								
Reliability Index	<.60	0.60–0.64	0.65–0.69	0.70–0.74	0.75–0.79	0.80–0.84	0.85–0.89	0.90–0.94	0.95–1.00
Percent Classification Agreement	< .001	< .001	.002	.009	.062	.188	.295	.234	.210
Tetrachoric Correlation	.016	.014	.036	.071	.089	.152	.248	.260	.114
Cohen’s Kappa	.410	.150	.165	.137	.072	.031	.017	.010	.009
$\hat{\gamma}_d$, Wang et al. (2015)	.017	.019	.052	.099	.178	.198	.206	.163	.069
$\hat{\mathrm{P}}_{cd}$, Johnson and Sinharay (2018)	.060	.076	.123	.140	.123	.126	.093	.077	.182

Figure 2 shows the relationship between the indices within linkage level in the form of a ternary diagram (West, 1982). In these plots, linkage levels where the three indices being compared have a similar value are in the central area. If one index is much larger than the other two, the linkage level shifts toward that corner of the figure. If one index is smaller than the other two, the linkage level shifts toward the side opposite of that index's corner (von Eynatten et al., 2002).

Figure 2

Similarity of Reliability Indices



Note. Each summary of the simulated retests was compared to the non-simulation approaches individually. Panel A: Similarity with the percent classification agreement. Panel B: Similarity with the tetrachoric correlation. Panel C: Similarity with the Cohen's kappa.

For all summaries of the simulated retests, the majority of linkage levels fell within the central area of the ternary diagrams, indicating a general level of agreement between the summary statistics and the non-simulation-based indices. However, within the central areas of Figure 2, there is only a weak relationship between each summary statistic and the other reliability indices. Table 4 shows the correlations between each simulation summary statistic and the other reliability indices, as well as the total number of linkage levels that fell outside of the central area of the ternary diagram. Overall, the correlations of the simulated retest summary statistics with the $\hat{\gamma}_d$ values from Wang et al. (2015) were larger than those with the $\hat{\mathrm{P}}_{cd}$ values from Johnson and Sinharay (2018). The correlations between the simulated retest summary statistics and $\hat{\gamma}_d$ were in the moderate to large range, whereas the correlations with $\hat{\mathrm{P}}_{cd}$ were in the small to moderate range (Cohen, 1988, 1992).

Table 4

Correlations Between Reliability Indices

Summary Statistic	$\hat{\gamma}_d$	$\hat{\mathrm{P}}_{cd}$	Linkage Levels Outside Central Area
Percent Classification Agreement	.22	.15	0
Tetrachoric Correlation	.55	.13	5
Cohen's Kappa	.53	.12	100

There were 100 linkage levels (7%) that fell outside of the central areas of Figure 2. These linkage levels are summarized in Table 5. The ten linkage levels flagged for the tetrachoric correlation (Figure 2B) were also flagged for the Cohen's kappa (Figure 2C). The linkage levels that were flagged for both the tetrachoric correlation and Cohen's kappa included two Successor level ELA linkage levels, four Precursor level science linkage levels, and four Target level science linkage levels. Notably, the majority of the flagged linkage levels were at the Target and Successor levels (*n* = 82; 82%), with the Successor level ELA linkage levels as the modal category (*n* = 31; 31%). The Target and Successor levels are tested less often than other linkage levels (Clark et al., 2019). Thus, the reliability indices for these linkage levels are based on a smaller sample size, as relatively few students in the resampling for the simulated retests would be assessed on these linkage levels compared to the precursor linkage levels. Additionally, of the 82 Target and Successor linkage levels that were flagged, 49 (60%) have a base rate of mastery either below .30 or above .70. As previously noted, Cohen's kappa can be overly conservative when there is a large class imbalance (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Combined with the smaller sample sizes, this likely explains the lower values observed for the Cohen's kappa, relative to the other reliability indices.

Table 5

Linkage Levels Outside the Central Area of Ternary Diagram, by Subject and Linkage Level

Linkage Level	English Language Arts	Mathematics	Science
Initial Precursor	0	0	2
Distal Precursor	0	0	—
Proximal Precursor	2	4	10
Target	6	6	13
Successor	31	26	—

Note. Science has three linkage levels: Initial, Precursor, and Target.

6. Discussion

The simulated retest approach is an innovative method for the estimation of reliability. However, despite the body of theoretical research supporting this method (e.g., Anozie & Junker, 2007; Sinharay & Johnson, 2019; Templin

& Bradshaw, 2013; Thompson et al., 2019), the simulation method has not previously been compared empirically to other notable DCM-based reliability indices. This report sought to compare the reliability estimates of DLM assessments from the simulated retests to reliability from non-simulation methods. The results from this study indicate that overall, the linkage level reliability summaries from the simulated retests generally agree with the non-simulation indices from Wang et al. (2015) and Johnson and Sinharay (2018). Relatively few linkage levels showed significant disagreement across the indices, and those that did were primarily linkage levels with small samples sizes and unbalanced class membership.

These findings indicate the summaries of classification consistency from the simulated retests are consistent with the reliability estimates derived from more traditional, non-simulation-based methods. However, the simulated retest method also allows for the reporting of results at aggregated levels of reporting (Thompson et al., 2019). Whereas the indices of Wang et al. (2015) and Johnson and Sinharay (2018) are limited to attribute-level consistency, the simulated retest method allows for the reporting of reliability for combinations of attributes. Cui et al. (2012) proposed a method for pattern-level consistency of DCMs, but the method also represents only a single level of reporting, as the proposed method does not extend down to individual attributes nor up to aggregations of patterns. Thus, the simulated retests offer the most flexibility in terms of what level of reliability can be reported, which is critical for DLM assessments and other operational programs that desire to provide both fine-grained reporting to inform instruction but also overall achievement to include in accountability models.

Finally, although this report focused on estimating the reliability of linkage levels within DLM assessments, the simulated retest methodology is generalizable. Given the flexibility to evaluate reliability at multiple levels of reporting and the consistency with traditional methods, the simulated retest methodology could be applied to any DCM-based assessment. However, future work should examine how test characteristics influence the results from the simulated retests to facilitate a better understanding of the conditions under which the simulation method performs best.

7. References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, *1*, 201–218. https://doi.org/10.1177/096228029200100205

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

Anozie, N., & Junker, B. (2007). *Investigating the utility of a conjunctive model in Q-matrix assessment using monthly student records in an online tutoring system* [Paper presentation]. National Council on Measurement in Education annual meeting, Chicago, IL.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, *27*, 3–23. https://doi.org/10.2307/3315487

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

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement butt low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*, 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Clark, A. K., Thompson, W. J., & Karvonen, M. (2019). *Instructionally embedded assessment: Patterns of use and outcomes* (Technical Report No. 19-01). University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/IE_Usage_Report_2018.pdf

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. https://doi.org/10.1037/h0026256

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

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

Dynamic Learning Maps Consortium. (2016a). *2014–2015 Technical Manual—Integrated Model*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_IM_2014-15.pdf

Dynamic Learning Maps Consortium. (2016b). *2014–2015 Technical Manual—Year-End Model*. University of Kansas, Center for Educational Testing and Evaluation. https: //dynamiclearningmaps.org/sites/default/files/documents/publication/Technical_Manual_YE_2014-15.pdf

Dynamic Learning Maps Consortium. (2017). *2015–2016 Technical Manual Update—Integrated Model*. University of Kansas, Center for Educational Testing and Evaluation. https://dynamiclearningmaps.org/sites/default/files/documents/publication/DLM_Technical_Manual_IM_2015-16.pdf

Dynamic Learning Maps Consortium. (2019). *2018–2019 Technical Manual Update—Integrated Model*. University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS). https://dynamiclearningmaps.org/sites/default/files/documents/publication/2018-2019_IM_Technical_Manual_Update.pdf

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement butt low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Hripcsak, G., & Heitjan, D. F. (2002). Measuring agreemen in medical informatics reliability studies. *Journal of Biomedical Informatics*, *35*, 99–110. https://doi.org/10.1016/S1532-0464(02)00500-2

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

Johnson, M. S., & Sinharay, S. (2019). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998619864550

Kirk, D. (1973). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, *38*, 259–268. https://doi.org/10.1007/BF02291118

O'Leary, S., Lund, M., Ytre-Hauge, T. J., Holm, S. R., Naess, K., Dalland, L. N., & McPhail, S. M. (2014). Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy*, *100*, 27–35. https://doi.org/10.1016/j.physio.2013.08.002

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

Pontius, R. G., Jr., & Millones, M. (2011). Death to kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*, 4407–4429. https://doi.org/10.1080/01431161.2011.552923

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

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

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

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

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

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

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

von Eynatten, H., Pawlowsky-Glahn, V., & Egozcue, J. J. (2002). Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams. *Mathematical Geology*, *34*, 249–257. https://doi.org/10.1023/A:1014826205533

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

Warrens, M. J. (2008). On association coefficients for 2 × 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*, 777–789. https://doi.org/10.1007/s11336-008-9070-3

West, D. R. E. (1982). *Ternary equilibrium diagrams* (2nd). Springer Netherlands. https://doi.org/10.1007/978-94-009-5910-1