Symposium on

Diagnostic Assessments: Moving from Theory to Practice

Title:

Technical Evidence for Diagnostic Assessments

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

ATLAS, University of Kansas

Author Note

Paper presented at the 2021 annual meeting of the National Council of Measurement in Education, virtual conference. Correspondence concerning this paper should be addressed to W. Jake Thompson, ATLAS, University of Kansas, 1122 West Campus Road, Lawrence, KS, 66045; jakethompson@ku.edu. Do not redistribute this paper without permission of the authors.

Abstract

Diagnostic classification models (DCMs) have grown in popularity over the past decade. However, their adoption in applied settings, especially operational assessment programs, has been minimal to slow. One potential barrier to adoption is the technical evidence recommended for all assessments in the *Standards for Educational and Psychological Testing*. Many of the methods widely used to provide evidence to meet these recommendations have implicit or explicit assumptions of a continuous unidimensional scale, such as those found in classical test theory and item response theory. In this paper, we describe how the use of a DCM impacts the type of technical evidence that should be provided for an assessment system, as well as methods for providing that evidence. An applied example from an operational assessment program that uses a DCM for reporting is provided, demonstrating how technical evidence can be provided for DCM-based assessments. We provide recommendations for other programs seeking to adopt a diagnostic assessment.

Keywords: diagnostic classification models, validity, reliability, fairness, DIF

Technical Evidence for Diagnostic Assessments

All operational assessment programs should provide technical evidence to support the claims and intended uses of the assessment. The *Standards for Educational and Psychological Testing* ("*Standards*" hereafter; American Educational Research Association [AERA] et al., 2014) describe best practices for documenting technical evidence for a wide range of tests, including educational assessment. The *Standards* identifies three foundations of any assessment that should be supported by evidence: validity, reliability, and fairness. Additionally, the *Standards* provide best practices for assessment operations (e.g., test development, standard setting, score reporting, etc.) and testing applications (e.g., psychological testing, educational assessment, etc.).

Although the *Standards* provide a wide breadth of recommendations for evidence to support the use of assessments, they are not comprehensive. For example, most of the recommendations for types of evidence that should be documented are based on premise of individual results being provided as a continuous scale-score. Continuous scale-scores are the person-level result for assessments that have been scaled with classical test theory (CTT; Lord & Novick, 1968) or item response theory (IRT; Birnbaum, 1968; Lord, 1953). Though CTT or IRT are used for many assessment systems, they are not used for all. Specifically, diagnostic classification models (DCMs; Rupp et al., 2010; Bradshaw, 2016) are a relatively recent alternative to traditional CTT and IRT approaches for assessment scaling.

Rather than estimating student-level results as a continuous scale, DCMs are inherently multivariate and assume a categorical underlying latent trait. In practice, this means that DCMs are able to provide fine-grained profiles of student achievement for a given set of knowledge, skills, and understandings (i.e., attributes). The categorical nature of the latent traits in a DCM-based assessment has implications for how technical evidence can be provided. Because, many of the standard methods for providing technical evidence assume a continuous latent trait, these methods are not necessarily applicable to assessments that use a DCM. Therefore, these methods must either be modified or substituted with a more appropriate method that is consistent with the construct of interest and measurement model.

In this paper we describe how aspects of the foundational technical evidence (i.e., validity, reliability, and fairness) may need to be adapted for diagnostic assessments. We provide a high-level overview of DCMs and discuss how unique characteristics of these models impact the type of evidence that is typically provided for validity, reliability, and fairness. An applied example using the Dynamic Learning Maps alternate assessment is then presented to demonstrate how the unique DCM considerations manifest in an operational assessment.

Diagnostic Classification Models

DCMs are a class of psychometric models that define a mastery profile for each individual on a pre-defined set of knowledge, skills, and understandings, referred to as attributes. The attributes in a DCM are categorical, and although they can consist of more than two categories, most applications of DCMs use dichotomous attributes (Rupp & Templin, 2008). Assuming binary attributes, the total number of mastery profiles is given as $C = 2^A$, where A is the total number of attributes in the DCM. Given a mastery profile for an individual, the probability of providing a correct response to an item is determined by the attributes that are required by the item. The relationships between attributes and items are defined in an I by A matrix, called a Q-matrix, where I is the number of items (Tatsuoka, 1983). Thus, the probability of individual r providing a response to an item is as shown in Equation 1.

$$P(\mathrm{X}_r = \mathrm{x}_r) = \sum_{c=1}^{C} \nu_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1 - \pi_{ic})^{1-x_{ir}} \qquad (1)$$

In Equation 1, ν_c represents the base rate of membership in class c, and π_{ic} is the probability of an individual in class c providing a correct response to item i. These are the parameters that are estimated for the DCM. In most cases, ν is left unstructured; that is, the base rate for each class if estimated directly without constraint (Rupp et al., 2010). In contrast, there are numerous ways that the π parameters can be estimated, depending on whether or not theory suggests attributes are able to compensate for each other in instances where an item measures multiple attributes. For example, we could define π using the deterministic-input, noisy-and-gate (DINA; Junker & Sijtsma, 2001); deterministic-input, noisy-or-gate (DINO; Templin & Henson, 2006); or loglinear cognitive diagnostic model (LCDM; Henson et al., 2009; Henson & Templin, 2019), just to name a few. Although the choice of DCM is a critical decision that should be driven by both cognitive theory and empirical evaluation, a discussion of the model selection process is beyond the scope of this paper (for an overview of these methods, see Chen et al., 2013; Sen & Bradshaw, 2017).

Once the model has been estimated, individuals receive results, or scores, in the form of a mastery profile. Typically, the results of assessments using DCMs with binary attributes are provided as a profile of dichotomous “master” or “non-master” decisions for each attribute; however, the raw probability of mastery for each attribute may also be reported (Bradshaw & Levy, 2019). Thus, in contrast to CTT and IRT methods where the focus of scoring is a single overall general ability score, DCMs are designed to provide fine-grained multivariate scores based on the attributes defined in the test design.

In summary, DCMs differ from CTT and IRT methods in two keys ways. First, when using DCMs, the focus is on placing individuals into categories, rather than locating them along a continuum. Second, DCMs are multivariate by design. Rather than a unidimensional scale as is common for most applications of CTT and IRT in operational assessment, DCM-based methods are intended to report scores for multiple latent traits simultaneously. These characteristics of DCMs have significant implications for the technical evidence that must be provided to support the use of DCM-based scores in an operational assessment. The following sections describe how these characteristics impact validity, reliability, and fairness evidence for DCMs.

Validity

Standard 1.0 of the *Standards* states that "clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA et al., 2014, p. 23). This includes providing evidence related to content, response process, internal structure, relation to other variables, and consequences (AERA et al., 2014). Across the three foundational areas, validity is the most similar between CTT, IRT, and DCMs.

Modern validity theory utilizes an argument-based approach to providing evidence to support the intended uses of test scores (Kane, 1992, 2002, 2006, 2009). This involves stating each claim of the assessment up front, and then providing evidence for each claim, and identifying places where there are gaps and additional evidence is needed. An effective way to frame this type of argument is through a theory of action (e.g., Bennett et al., 2011; Chalhoub-Deville, 2016; Perie & Forte, 2011). A theory of action includes all of the claims made by an assessment system and describes how the claims are connected to each other through a logical chain of reasoning. Regardless of the psychometric model used to score the assessment, some claims would be common (e.g., items are correctly aligned, students are able to interact with assessment system, etc.).

On the other hand, some claims may be specific to the use of a DCM. For example, because DCM results produce a fine-grained mastery profile, we might expect for there to be claims related to the utility of attribute-level scores for instructional planning or other uses. Relatedly, we might expect an assessment using a DCM to include claims related to the mastery classifications. Similar to how assessments using CTT and IRT would include claims about the accuracy of student scores, DCM-based assessments should include evidence to support the claim that reported classifications are accurate. Additionally, if the DCM-based assessment is used to make summative statements about overall student achievement, there will likely be a claim about student results across all tested attributes representing overall performance. In a way, DCM-based assessments are the inverse of CTT- and IRT-based assessments. For CTT- and IRT-based assessments, the primary score is the overall score, which in some instances may be broken down in subscores, for which additional evidence must be provided (Feinberg & Wainer, 2014; Sinharay et al., 2011). In contrast, the primary "score" for DCM-based assessments is a profile of mastery on more fine-grained attributes (analogous to subscores). Thus, any summative judgements would require additional evidence to support the aggregation or cohesion of the individual attribute results.

Overall, the framework for constructing a validity argument for a DCM-based assessment is not that different from that used for CTT- or IRT-based assessments. Instead, the differences are mainly in the types of claims that might be included and the types of evidence that may be required for each claim.

Reliability

Reliability refers to the precision or consistency of test scores. In the *Standards*, it is recommended that "appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (Standard 2.0; AERA et al., 2014, p. 42). That is, for each reported score, there should be a level of precision documented to indicate the amount of error that may be present in the estimated score. In addition to the reliability of test scores, the *Standards* recommend that test documentation include the decision consistency for any classifications that are made using the scores (Standard 2.16; AERA et al., 2014, p. 46). This includes performance levels on academic achievements tests. For example, if an assessment reports both an overall score and a performance level, documentation should include an estimate of reliability for the overall test score, as well as the decision consistency of the performance level classification.

In assessments scaled using CTT, score reliability is typically reported using a reliability index such as the Guttman-Cronbach alpha (Cronbach, 1951; Guttman, 1945), coefficient omega (Jöreskog, 1971; McDonald, 1999), or the greatest lower bound (Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977). These measures all attempt to quantify the amount of variance in the total score that is due to the underlying "true" score relative to the amount of variance in the total score that is due to error. Values close to 1.0 indicate a high level of reliability (i.e., almost all the variance in observed scores is due to the true score), and values close to 0.0 indicate poor reliability. Once the score reliability has been estimated, the decision consistency of the test can be assessed using many methods. The most popular is the Livingston and Lewis (1995) method, which uses an effective test length estimate and an assumed beta-binomial model to estimate both the decision consistency and accuracy for a CTT-based classification.

Because assessments scored with IRT do not assume that all items provide the same amount of information, alternative methods have been developed for score reliability and decision consistency. In IRT-based assessments, items provide different amounts of information for different values of the latent trait. Thus, the precision of the latent trait score is conditional on the score itself. This results in what is commonly referred to as the conditional standard error of measurement (Nicewander, 2019). Thus, there is not a single reliability index as is common for assessments using CTT, but rather many estimates, one for each value of the latent score or scale score (Kolen et al., 1996). The conditional observed score distribution can then be used to calculate the decision consistency for achievement levels, as described by Lee (2010). Like the CTT methods described above, the IRT-based reliability methods also assume a continuous unidimensional trait.

Because results from DCM-based assessments are neither unidimensional nor continuous, the reliability indices from CTT and test information functions from IRT are inappropriate. However, there are many ways reliability can be assessed for discrete attributes. A review of notable reliability methods for DCMs is available from Sinharay & Johnson (2019). In general, the choice of reliability method is largely dependent on how results of the assessment are reported. If results are reported as probability of mastery, then the reliability should be reported at the precision of the estimated probability (e.g., Johnson & Sinharay, 2020; Templin & Bradshaw, 2013).

Alternatively, results reported as classification decisions should report the classification consistency. Classification consistency can be reported at the pattern level (e.g., Cui et al., 2012) or at the attribute level (e.g., Johnson & Sinharay, 2018; Wang et al., 2015). Alternatively, Thompson et al. (2019) proposed a simulation-based approach to estimating classification consistency that allows for reporting at multiple levels of reporting, including an overall achievement level, if desired. Thus, although the reliability methods for CTT- and IRT-based assessments are not appropriate for DCM-based assessments, there are well established methods for evaluating mastery probabilities and/or classifications.

Fairness

The topic of fairness is a broad concept that encompasses all aspects of the test development process from test design to score interpretations by end users. This is realized in the *Standards*, which state that "all steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended use for all examinees in the intended population" (Standard 3.0; AERA et al., 2014, p. 63). Due to the wide range of topics that fall under the umbrella of fairness, in this paper we focus on one aspect that is relevant for psychometrics and is typically evaluated for operational assessment programs: differential item function (DIF).

DIF refers to the functioning of items across groups of individuals, usually demographic such as sex, race, or age group (Camilli, 2006; Holland & Wainer, 1993), after holding achievement constant. This presents a critical threat to test fairness, as tests should not be affected by construct-irrelevant factors (e.g., Standard 3.2 and Standard 3.6; AERA et al., 2014). In practice DIF is assessed by evaluating whether or not group membership is a significant predictor of item performance, after accounting for overall ability (the matching variable). There are many methods that can be used to determine whether or not group membership is "significant." Two of the most popular are the Mantel-Haenszel method (Mantel & Haenszel, 1959) and logistic regression (Swaminathan & Rogers, 1990). Both methods attempt to compare, for each value of the matching variable, how differently two or more groups perform. When using CTT or IRT, the matching variable is usually the total sum score or the estimate latent trait, θ, respectively.

When using a DCM, both the Mantel-Haenszel and logistic regression methods can still be used, but special attention should be paid to the matching variable (Qiu et al., 2019). One option is to use the full mastery profile as the matching variable. Assuming binary attributes, there would be 2^A possible profiles. This quickly creates a large number of classes, many of which likely have small sample sizes, making estimation of the DIF model challenging. Alternatively, one could collapse profiles that the DCM specifies should respond the same to an item. For example, take an assessment that measures three attributes in total. If an item measures only the first attribute, then individuals in profiles [0,0,0], [0,1,0], [0,0,1], and [0,1,1] should have the same probability of providing a correct response because attribute one is 0 in all of these profiles, and attributes two and three are irrelevant to performance on this item. Collapsing profiles can help keep sample sizes from getting too small; however, this requires determining the matching variable for each item individually and will be less helpful for items measuring multiple attributes.

Additionally, an aggregation of the individual attributes (e.g., total attributes mastered) could be used as the matching variable. If the assessment includes many attributes, then an aggregated score begins to look similar to the CTT total score, and the matching variable no longer has to be treated as categorical. However, this could contradict model expectations. Continuing the previous example, individuals with profiles [0,1,1] and [1,0,1] have both mastered two attributes, but because the item measures only the first attribute, the model expects individuals within these profiles to respond differently, even though they have all mastered two total attributes. Thus, special consideration must be given to how the matching variable is defined for DIF analyses in a DCM-based assessment.

Technical Evidence in Practice: Dynamic Learning Maps

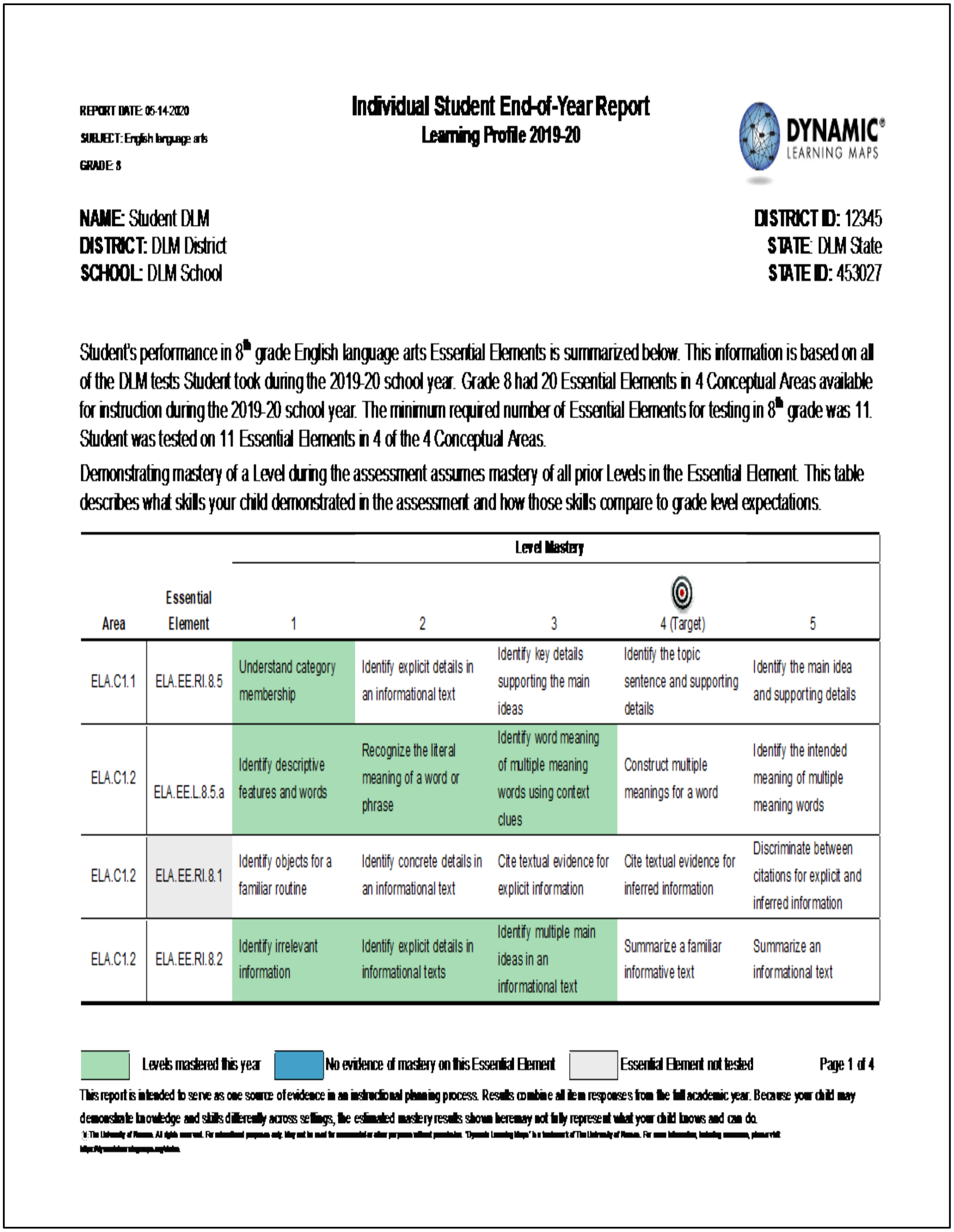
Dynamic Learning Maps (DLM) alternate assessments are administered in over 20 states to students with the most significant cognitive disabilities. The DLM assessments evaluate student learning and achievement in English language arts (ELA), mathematics, and science. Assessments are scored using DCMs, making the DLM system the first large-scale application of DCM-based assessments for statewide accountability purposes. In the DLM assessments, the alternate content standards, or Essential Elements, are specific statements of knowledge, skills, and understandings. To ensure that all students are able to access the academic content, each Essential Element is associated with five linkage levels: three precursor levels, one target level aligned to the Essential Element, and one level that extends beyond the grade-level expectation.[1]

The linkage level is the unit of scoring for DLM assessments, and thus represent the attributes in the DCM. Linkage levels are assessed using testlets, which consist of three to five items and are centered around an engagement activity. For each assessed linkage level, students receive a “mastered” / “not mastered” decision reported in a fine-grained learning profile (Figure 1). In addition to the individual mastery decisions, summative achievement is also reported as the number of linkage levels mastered within each conceptual area, or content strand, as well as an overall performance level for each subject (Figure 2). The use of a DCM for scoring, and providing results at multiple levels of reporting, has implications for the evidence needed to support the claims of the assessment. In the following sections, we describe how we provide technical evidence for DLM assessments in the areas of validity, reliability, and DIF.

[1] Essential Elements for the DLM science assessment include three linkage levels: Initial, Precursor, and Target.

Figure 1

Learning Profile Showing Fine-Grained Mastery Results for DLM Assessments



REPORT DATE: 05-14-2020
SUBJECT: English language arts
GRADE: 8

Individual Student End-of-Year Report
Learning Profile 2019-20

DYNAMIC® LEARNING MAPS

NAME: Student DLM
DISTRICT: DLM District
SCHOOL: DLM School

DISTRICT ID: 12345
STATE: DLM State
STATE ID: 453027

Student's performance in 8th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2019-20 school year. Grade 8 had 20 Essential Elements in 4 Conceptual Areas available for instruction during the 2019-20 school year. The minimum required number of Essential Elements for testing in 8th grade was 11. Student was tested on 11 Essential Elements in 4 of the 4 Conceptual Areas.

Demonstrating mastery of a Level during the assessment assumes mastery of all prior Levels in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

		Level Mastery				
Area	Essential Element	1	2	3	4 (Target)	5
ELA.C1.1	ELA.EE.RI.8.5	Understand category membership	Identify explicit details in an informational text	Identify key details supporting the main ideas	Identify the topic sentence and supporting details	Identify the main idea and supporting details
ELA.C1.2	ELA.EE.L.8.5.a	Identify descriptive features and words	Recognize the literal meaning of a word or phrase	Identify word meaning of multiple meaning words using context clues	Construct multiple meanings for a word	Identify the intended meaning of multiple meaning words
ELA.C1.2	ELA.EE.RI.8.1	Identify objects for a familiar routine	Identify concrete details in an informational text	Cite textual evidence for explicit information	Cite textual evidence for inferred information	Discriminate between citations for explicit and inferred information
ELA.C1.2	ELA.EE.RI.8.2	Identify irrelevant information	Identify explicit details in informational texts	Identify multiple main ideas in an informational text	Summarize a familiar informative text	Summarize an informational text

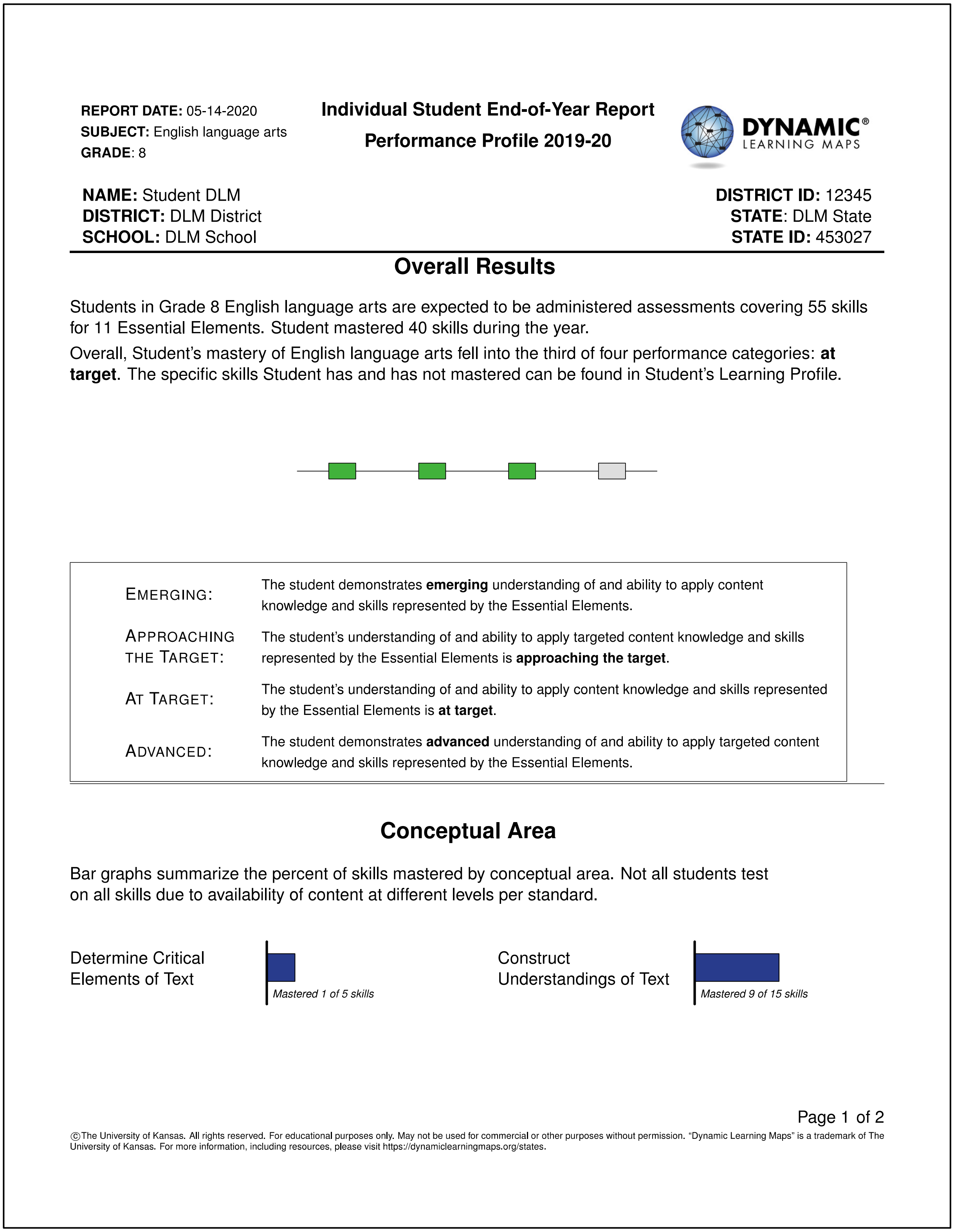
🟩 Levels mastered this year 🟦 No evidence of mastery on this Essential Element ⬜ Essential Element not tested Page 1 of 4

This report is intended to serve as one source of evidence in an instructional planning process. Results combine all item responses from the full academic year. Because your child may demonstrate knowledge and skills differently across settings, the estimated mastery results shown here may not fully represent what your child knows and can do.

©The University of Kansas. All rights reserved. For educational purposes only. May not be used for commercial or other purposes without permission. "Dynamic Learning Maps" is a trademark of The University of Kansas. For more information, including resources, please visit https://dynamiclearningmaps.org/states.

Figure 2

Performance Profile Showing Overall Achievement for DLM Assessments



REPORT DATE: 05-14-2020
SUBJECT: English language arts
GRADE: 8

Individual Student End-of-Year Report
Performance Profile 2019-20

DYNAMIC® LEARNING MAPS

NAME: Student DLM
DISTRICT: DLM District
SCHOOL: DLM School

DISTRICT ID: 12345
STATE: DLM State
STATE ID: 453027

Overall Results

Students in Grade 8 English language arts are expected to be administered assessments covering 55 skills for 11 Essential Elements. Student mastered 40 skills during the year.

Overall, Student's mastery of English language arts fell into the third of four performance categories: **at target**. The specific skills Student has and has not mastered can be found in Student's Learning Profile.

🟩 🟩 🟩 ⬜

EMERGING:	The student demonstrates **emerging** understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
APPROACHING THE TARGET:	The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is **approaching the target**.
AT TARGET:	The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is **at target**.
ADVANCED:	The student demonstrates **advanced** understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

Conceptual Area

Bar graphs summarize the percent of skills mastered by conceptual area. Not all students test on all skills due to availability of content at different levels per standard.

Determine Critical Elements of Text — *Mastered 1 of 5 skills*

Construct Understandings of Text — *Mastered 9 of 15 skills*

Page 1 of 2

Validity Argument for DLM Assessments

Validity for the DLM assessments is evaluated in the context of a theory of action, as described by Clark & Karvonen (2020). In addition to claims that would be common to most large-scale assessment systems, the DLM theory of action includes additional claims specific to the population of students taking the DLM assessment, as well as the use of a DCM for the scoring model. The DLM theory of action organizes 17 claims into 4 categories: design, delivery, scoring, and outcomes. The full theory of action for DLM assessments is shown in Figure 3. The use of a DCM has indirect implications for many of the claims in the theory of action. For example, the use of a DCM informs the test development process and design, which in turn impacts assessment delivery. Thus, even though these types of assessment design and delivery claims would likely be present for non-DCM-based assessments, the use of a DCM may or may not impact how evidence is provided. In this paper, we will focus on claims that are directly impacted by the choice of a DCM. In the DLM theory of action, these are the three claims under the scoring category (Figure 3):

K. Mastery results indicate what students know and can do.

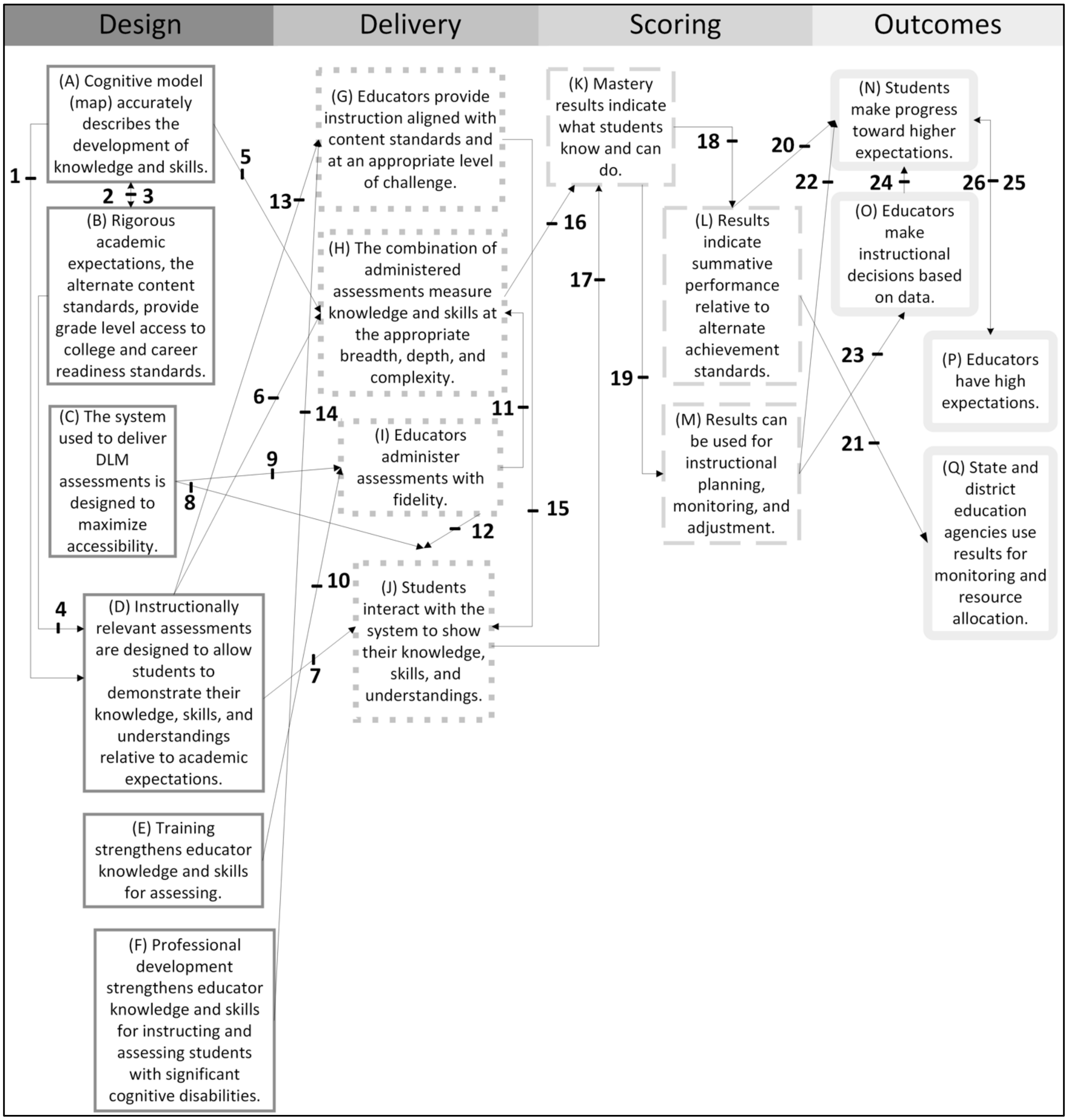
L. Results indicate summative performance relative to alternate achievement standards.

M. Results can be used for instructional planning, monitoring, and adjustment.

The first claim, Claim K, is directly tied to the mastery classifications that are unique to DCMs. That is, evidence should be provided that the mastery classifications are accurately representing what knowledge, skills, and understandings the student has mastered.

Figure 3

Theory of Action for DLM Assessments


Design
Delivery
Scoring
Outcomes
(A) Cognitive model (map) accurately describes the development of knowledge and skills.
(B) Rigorous academic expectations, the alternate content standards, provide grade level access to college and career readiness standards.
(C) The system used to deliver DLM assessments is designed to maximize accessibility.
(D) Instructionally relevant assessments are designed to allow students to demonstrate their knowledge, skills, and understandings relative to academic expectations.
(E) Training strengthens educator knowledge and skills for assessing.
(F) Professional development strengthens educator knowledge and skills for instructing and assessing students with significant cognitive disabilities.
(G) Educators provide instruction aligned with content standards and at an appropriate level of challenge.
(H) The combination of administered assessments measure knowledge and skills at the appropriate breadth, depth, and complexity.
(I) Educators administer assessments with fidelity.
(J) Students interact with the system to show their knowledge, skills, and understandings.
(K) Mastery results indicate what students know and can do.
(L) Results indicate summative performance relative to alternate achievement standards.
(M) Results can be used for instructional planning, monitoring, and adjustment.
(N) Students make progress toward higher expectations.
(O) Educators make instructional decisions based on data.
(P) Educators have high expectations.
(Q) State and district education agencies use results for monitoring and resource allocation.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

Claim L is not necessarily unique to DCM-based assessments. We would expect many assessment systems to include a claim regarding summative performance. However, because the summative achievement level is not directly estimated by the DCM, additional evidence is needed to support the aggregation of separate linkage level mastery classifications into an overall summative indicator of achievement.

Finally, Claim M is related to fine-grained nature of DCM-based reporting. Because results are reporting each individual linkage level, representing specific knowledge, skills, and understandings, we intend for assessment results to be used in ways that may not be possible when only a single scale score is reported. Thus, evidence should be provided to demonstrate that the classifications are accurate (Claim K), but also that the classifications are reported out in a manner that is understandable and actionable for a variety of stakeholders.

In summary, the DLM theory of action provides the framework to for the validity argument in support of the intended uses of the DLM assessments. Although the use of a theory of action is not unique to DCM-based assessments, some of the claims and required evidence are. Therefore, it is important to fully explore the implications of using of DCM and the evidence necessary to evaluate claims impacted by the selection of a DCM-based scoring model when developing the validity argument. Although it is beyond the scope of this paper to describe specific evidence for each claim, this evidence is available in the DLM technical manual (DLM Consortium, 2016) and in the subsequent annual updates to the technical manual (e.g., DLM Consortium, 2017, 2020).

Reliability for DLM Assessments

Reliability for DLM assessments is evaluated using a simulated re-test design described by Thompson et al. (2019) and in Chapter 8 of DLM Consortium (2017). At a high level, the simulation process adheres to the process:

1. Draw with replacement a student record from the operational data set.
2. Using calibrated model parameters and estimated mastery of tested linkage levels,

simulate item responses from a hypothetical second test administration.

3. Score the simulated responses using operational scoring rules, including any aggregated performance indicators (e.g., conceptual areas or overall performance level).
4. Repeat steps 1–3 for 2,000,000 simulated students.

The mastery determinations and aggregations from the simulated re-tests are then compared to corresponding results from the observed test to assess the consistency in the scores. For DLM assessments, consistency is estimated using Cohen's κ (Cohen, 1960, 1968), tetrachoric and polychoric correlations (Bonett & Price, 2005), and percent classification agreement. These measures were chosen because in addition to being widely used and thoroughly vetted in the research literature, these methods provide complementary information, with the strengths and weaknesses of each method balancing each other (Thompson, 2020).

The simulation method is useful because it allows for the evaluation of reliability at multiple levels of reporting. Most of the DCM reliability methods are focused on individual attributes (Sinharay & Johnson, 2019). However, these methods do not allow for the evaluation of reliability for assessments reporting aggregated summaries of attributes mastery (e.g., achievement levels), and which should describe the reliability of the aggregated scores in technical documentation (Standard 2.3, AERA et al., 2014). At the attribute level, the simulation methodology compares favorably to non-simulation approaches. Thompson (2020) compared the attribute-level classification consistency indices from the simulation method (i.e., tetrachoric correlation, Cohen's κ, and percent correct classification) to the indices proposed by Wang et al. (2015) and Johnson & Sinharay (2018). The comparisons were favorable, with a high-level of agreement across the simulation and non-simulation indices. Thus, the reliability simulation method used for DLM assessments is able to provide attribute-level classification consistency measures that are themselves consistent with existing indices, as well as provide reliability evidence for additional levels of aggregated reporting.

DIF for DLM Assessments

As with many operational tests, DIF for the DLM assessments is evaluated using the logistic regression procedure. A complete description of the DIF method is included in Chapter 9 of the annual technical manual update (DLM Consortium, 2020). In summary, DIF is evaluated for both sex and racial groups. In total three models are estimated. The first includes only the matching variable as a predictor of item performance. The second model adds the grouping variable (i.e., gender or race) to evaluate for uniform DIF. The third model adds an interaction term between the matching variable and grouping variable to evaluate for non-uniform DIF. For each model comparison the Nagelkerke (1991) pseudo-R^2 is calculated, which is then categorized as a negligible, moderate, or large effect using the criteria of Zumbo and Thomas (1997) and Jodoin and Gierl (2001). Items identified with a moderate or large effect size are reviewed by test development teams.

In the DLM DIF models, the matching variable is the total number of linkage levels mastered within each subject. Thus, the matching variable used is a semi-continuous measure, similar to what would be used in a CTT context. The total number of linkage levels available for testing in a given grade and subject ranges from 27 to 100 unique linkage levels, due to differences in the scope of blueprints (i.e., how many standards are available). Thus, there are enough attributes to create a reasonable range for the matching variable. Hoover et al. (2020) compared DIF identification for DLM assessments when different matching variables were used. Specifically, the total number of linkage levels mastered was compared to a binary matching variable indicating mastery of the linkage level the items measures. Overall, the choice of the matching variable had minimal impact on whether an item was identified as having non-negligible DIF. However, because students typically only take three to five items for a single linkage level, the binary matching variable is more susceptible to bias. For example, if one item has DIF, that would represent 20% of the total items taken by a student. Thus, the mastery classification could be biased, in turn biasing the DIF analysis. Further, because there are only three to five items scale purification is not feasible (Qiu et al., 2019). Thus the aggregated matching variable is used for the operational analyses, as this measure is more robust to possible contamination, and does not meaningful impact the identification of DIF.

Discussion

DCMs are a powerful tool for understanding student learning. Results from DCM-based assessments are more fine-grained than the overall total or scale scores that are provided by CTT- and IRT-based assessments. Thus, DCM-based assessments can provide results that are more actionable (Clark et al., 2018). Mastery results can be used to inform subsequent instructional plans on individual standards or within content strands, and inform instructional groups and IEP goals. Additionally, DCMs can be used to better understand the acquisition of knowledge by evaluating theorized cognitive models and learning progressions (e.g., Templin & Bradshaw, 2014; Thompson & Nash, 2019) and measure student progress (e.g., Madison & Bradshaw, 2018; Zhan, 2021). Despite these benefits, adoption of DCMs in operational assessments settings has been slow.

There are many reasons why the use of DCMs has been limited in operational settings. Switching an assessment to a DCM-based scoring model is not as straightforward as swapping out different IRT models might be. In all assessment there is an interplay and dependency between the scoring model and assessment design. Thus, moving to a new model requires more than changing a scoring algorithm. Rather, the assessment theory of action would likely be impacted, and many of aspects of the assessment would need to be updated. Because DCMs are not widely used, how to implement these changes, and the necessary evidence to support the changes, may not always have clear answers. Additionally, there are policy and political context in which stakeholders may not yet be ready to move away from the more traditional CTT- and IRT-based approaches. In particular, switching scoring models would have implications for student growth and state accountability systems that are already in place. Thus, in addition to reflecting on the *Standards* and best practices in the literature, program staff should engage stakeholders early in the process to communicate the benefits of DCMs and achieve buy-in.

Among the many considerations when contemplating a switch to a DCM-based assessment is the ability to provide high-quality technical evidence. Many of the methods most commonly used to provide technical evidence have an implicit or explicit assumption that the reported scores are continuous and unidimensional. This is not the case for DCM-based assessments. However, a lack of a continuous scale score does not mean that technical evidence cannot be provided. Some existing methods can be adopted for use with DCMs, some evidence will require newer methods that have not had enough exposure to raise awareness of their existence, and yet other evidence requires further exploration and research. For example, most methods for evaluating the model fit of DCMs relies on limited information indices that are not able to capture the full complexities of the observed data (e.g., Chen et al., 2018). Using posterior predictive model checks from Bayesian estimation methods have been proposed to more robustly evaluate model fit (e.g., Park et al., 2015; Thompson, 2019); however, more research is needed in this area to understand the contexts in which posterior predictive checks are the most effective.

There is also more work to be done aggregating and summarizing performance across many attributes. For the DLM assessments, the total number of attributes mastered is summed, and then grouped into performance levels using cut scores (Clark et al., 2017). Alternatively, a higher-order latent trait model could be employed (e.g., de la Torre & Douglas, 2004, 2008), which estimates a broad latent trait similar to IRT in addition to the fine-grained mastery information. Although the higher-order latent trait is an interesting alternative, there has been little research to understand the circumstances under which this method would be preferable in an operational setting. Additional work would also be needed to evaluate what evidence would be needed to support the use of the IRT-like higher-order trait for summative reporting.

In this paper, we described how evidence of validity, reliability, and fairness are impacted by the choice to score an assessment using a DCM. We used the DLM assessments as an applied example where DCM-based evidence has been successful. The methods described in this paper have been reported in technical documentation for the assessment. Additionally, technical evidence used to support the use of DCMs in the DLM assessment system has been used by states to successfully meet the relevant peer review requirements (e.g., U.S. Department of Education, 2018), demonstrating the strength and rigor of the evidence for a DCM-based assessment. We hope that by describing the ways in which technical evidence can be provided for DCM-based assessments, we will encourage other organizations to consider how DCMs may benefit their stakeholders, without viewing the evidentiary requirements as prohibitive.

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Bennet, R. E., Kane, M., & Bridgeman, B. (2011, February 10–11). *Theory of action and validity argument in the context of through-course summative assessment* [Conference session]. Invitational Research Symposium on Through-Course Summative Assessments, Atlanta, GA, United States. https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Bennett_Kane_Bridgeman.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392–479). Addison-Wesley.

Bonett, D. G., & Price, R. M. (2005). Inferential Methods for the Tetrachoric Correlation Coefficient. *Journal of Educational and Behavioral Statistics*, *30*(2), 213–225. https://doi.org/10.3102/10769986030002213

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

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

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 220–256). American Council on Education/Praeger.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing, 33*(4), 453–472. https://doi.org/10.1177/0265532215593312

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling: Relative and absolute fit evaluation in CDM. *Journal Educational Measurement, 50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

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

Clark, A. K., & Karvonen, M. (2020). Constructing and evaluating a validation argument for a next-generation alternate assessment. *Educational Assessment, 25*(1), 47–64. https://doi.org/10.1080/10627197.2019.1702463

Clark, A. K., & Karvonen, M., Swinburne Romine, R., & Kingston, N. M. (2018, April 12–16). *Teacher use of score reports for instructional decision making: Preliminary findings* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY. https://dynamiclearningmaps.org/sites/default/files/documents/presentations/NCME2018_score%20report%20use.pdf

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological*

Measurement, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

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

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

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

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595–624. https://doi.org/10.1007/s11336-008-9063-2

Deng, N., & Hambleton, R. K. (2013). Evaluating CTT- and IRT-based single-administration estimates of classification consistency and accuracy. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & Woods, C. M. (Eds.), *New developments in quantitative psychology* (pp. 235–250). Springer. https://doi.org/10.1007/978-1-4614-9348-8_15

Dynamic Learning Maps Consortium. (2016). *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. (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. (2020). *2019–2020 technical manual update—Instructionally embedded model*. University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://2020-ie-techmanual.dynamiclearningmaps.org/

Feinberg, R. A., & Wainer, H. (2014). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational Measurement: Issues and Practice*, *33*(3), 47–54. https://doi.org/10.1111/emip.12037

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*. SAGE Publications.

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

Henson, R., & Templin, J. L. (2019). Loglinear cognitive diagnostic model (LCDM). In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 171–185). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_8

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Lawrence Erlbaum.

Hoover, J. C., Thompson, W. J., Clark, A. K., & Nash, B. (2020, April 16–20). *Diagnostic assessment: DIF detection from binary and aggregate mastery classifications* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San

Francisco, CA. (Conference canceled)

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogenous items: I: Algebraic lower bounds. *Psychometrika, 42*(4), 567–578. https://doi.org/10.1007/BF02295979

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power raters using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

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

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

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133. https://doi.org/10.1007/BF02291393

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

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. T. (2009).Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 39–64). Information Age Publishing.

Kolen, M. J., Lingjia, Z., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129–140. https://doi.org/10.1111/j.1745-3984.1996.tb00485.x

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1–17. https://doi.org/10.1111/j.1745-3984.2009.00096.x

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*(4), 517–548. https://doi.org/10.1177/001316445301300401

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

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

McDonald, R. P. (1999). *Test theory: A unified approach.* Erlbaum.

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Nicewander, W. A. (2019). Conditional precision of measurement for test scores: Are conditional standard errors sufficient? *Educational and Psychological Measurement, 79*(1), 5–18. https://doi.org/10.1177/0013164418758373

Park, J. Y., Johnson, M. S., & Lee, Y-S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education, 2*(3-4), 244–264. https://doi.org/10.1504/IJQRE.2015.071738

Perie, M., & Forte, E. (2012). Developing a validity argument for assessing students in the margin. In M. Russell & M. Kavanaugh (Eds.), *Assessing students in the margin: Challenges, strategies, and techniques* (pp. 335–378). Information Age Publishing.

Qiu, X.-L., Li, X., & Wang, W.-C. (2019). Differential item functioning in diagnostic classification models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 379–393). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_18

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262. https://doi.org/10.1080/15366360802490866

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

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347–362. https://doi.org/10.1177/01466219922031464

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement, 41*(6), 422–438. https://doi.org/10.1177/0146621617695521

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscores lack validity? Don't blame the messenger. *Educational and Psychological Measurement*, *71*(5), 789–797. https://doi.org/10.1177/0013164410391782

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 International Publishing. https://doi.org/10.1007/978-3-030-05584-4_17

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

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

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Thompson, W. J. (2019). *Bayesian psychometrics for diagnostic assessments: A proof of concept* (Research Report No. 19-01). University of Kansas; Accessible Teaching, Learning, and Assessment Systems. https://doi.org/10.35542/osf.io/jzqs8

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. https://dynamiclearningmaps.org/sites/default/files/documents/publication/Reliability_Comparison.pdf

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

Thompson, W. J., & Nash, B. (2019, April 4–8). Empirical methods for evaluating maps. In M. Karvonen (Chair), *Beyond learning progressions: Maps as assessment architecture* [Symposium]. National Council on Measurement in Education Annual Meeting, Toronto, Canada. https://bit.ly/maps-ncme19

United States Department of Education. (2018). [Letter Joy Hofmeister, Oklahoma State Department of Education]. Retrieved from https://www2.ed.gov/admins/lead/account/nclbfinalassess/ok3.pdf

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

https://doi.org/10.1111/jedm.12096

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogenous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42*(4), 579–591. https://doi.org/10.1007/BF02295980

Zhan, P. (2021). Refined learning tracking with a longitudinal probabilistic diagnostic model. *Educational Measurement: Issues and Practice*, *40*(1), 44–58. https://doi.org/10.1111/emip.12397

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* [Working Paper]. University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.