

Teacher Assessment Literacy: Implications for Diagnostic Assessment Systems

Dr. Amy K. Clark*
ATLAS
University of Kansas
1122 W Campus Road, Lawrence, KS 66045
akclark@ku.edu
ORCID: 0000-0002-5804-8336
Twitter: @atlas4learning

Dr. Brooke Nash
ATLAS
University of Kansas
1122 W Campus Road, Lawrence, KS 66045
bnash@ku.edu
ORCID: 0000-0001-9858-7062

Dr. Meagan Karvonen
ATLAS
University of Kansas
1122 West Campus Road, Lawrence, KS, 6604
karvonen@ku.edu
ORCID: 0000-0003-2071-2673

Author Note

Correspondence concerning this manuscript should be addressed to Amy K. Clark, ATLAS, University of Kansas, 1122 W. Campus Road, Lawrence, KS, 66045. Email: akclark@ku.edu

The authors would like to acknowledge Jennifer Burnes, Brianna Beitling, and Elizabeth Kavitsky for their data analysis contributions.

This is an Original Manuscript of an article published by Taylor & Francis in Applied Measurement in Education on February 1, 2022, available at <https://www.tandfonline.com/doi/full/10.1080/08957347.2022.2034823>.

Abstract

Assessments scored with diagnostic models are increasingly popular because they provide fine-grained information about student achievement. Because of differences in how diagnostic assessments are scored and how results are used, the information teachers must know to interpret and use results may differ from concepts traditionally included in assessment literacy trainings for assessments that produce a raw or scale score. In this study, we connect assessment literacy and score reporting literature to understand teachers' assessment literacy in a diagnostic assessment context as demonstrated by responses to focus groups and surveys. Results summarize teachers' descriptions of fundamental diagnostic assessment concepts, understanding of the diagnostic assessment and results produced, and how diagnostic assessment results influence their instructional decision-making. Teachers understood how to use results and were comfortable using the term *mastery* when interpreting score report contents and planning next instruction. However, teachers were unsure how mastery was calculated and some misinterpreted mastery as representing a percent correct rather than a probability value. We share implications for others implementing large-scale diagnostic assessments or designing score reports for these systems.

Keywords: assessment literacy, diagnostic assessments, score reporting, diagnostic modeling, instructional decision-making

Teacher Assessment Literacy: Implications for Diagnostic Assessment Systems

Building teachers' assessment literacy has been a focus for decades (e.g., Stiggins, 1991) to help support interpretation and use of assessment results. Various trainings and materials promote teachers' familiarity with measurement concepts (e.g., U.S. Department of Education, 2015). However, these resources are typically designed from a traditional classical test theory or item response theory approach to measurement. With the rise of innovative assessments (e.g., Every Student Succeeds Act Innovative Assessment Demonstration Authority [U.S. Department of Education, 2020]) and increased literature demonstrating the technical adequacy and utility of diagnostic modeling (e.g., Bradshaw, 2017), existing resources to support assessment literacy may not pertain to all assessment contexts. While innovative approaches to assessment may have potential to better support teachers' classroom instruction, these benefits are only realized to the extent that teachers know what the assessment results mean and how to use them.

Recent research demonstrates many benefits of assessments scored with diagnostic classification models (referred to as diagnostic assessments in this paper) over traditional measurement models. These assessments can produce reliable mastery classifications with fewer items than traditional assessments (Templin & Bradshaw, 2013). Diagnostic assessment score reports are delivered as mastery profiles that summarize the specific skills students demonstrated (Feldberg & Bradshaw, 2019; Karvonen et al., 2019). The design of fine-grained mastery reports supports teachers in subsequent instructional decision-making (Clark et al., 2018).

For teachers to use mastery profiles as intended, they should understand their contents and how to use the results to make subsequent instructional decisions. In other words, teachers should have assessment knowledge that is specific to diagnostic systems. As defined by Popham (2011), "Assessment literacy consists of individuals' *understandings* of the *fundamental*

assessment concepts and procedures deemed likely to influence educational decisions” (p. 267; emphasis in original). Because diagnostic assessment concepts and intended uses of results differ from more traditional, raw and scale score-based assessments, teachers using diagnostic assessments need a different type and breadth of assessment literacy than is required for traditional assessments. To this end, we used focus group and survey responses to understand teachers’ assessment literacy in the context of an operational diagnostic assessment system.

Supporting and Evaluating Assessment Literacy

Existing literature describes methods for supporting and evaluating teachers’ assessment literacy. A range of materials have been developed to support educators’ understanding of measurement concepts and how to use assessment results. Assessment literacy standards are available in many countries (e.g., The Standards for Teacher Competence in Educational Assessment of Students; American Federation of Teachers et al., 1990) and encompass many areas, including assessment purposes and processes, measurement theory, communicating about results, and supporting teachers’ competency (DeLuca et al., 2016b). Education agencies and other organizations make assessment literacy materials, such as professional development modules and descriptive documents, and credentialing available to instruct educators on key assessment concepts (e.g., Illinois State Board of Education, 2015; Kansas State Department of Education, 2019; Michigan Assessment Consortium, 2020; Wisconsin Department of Public Instruction, 2019).

Researchers have developed ways to evaluate teachers’ assessment literacy (e.g., DeLuca et al., 2016a; Lian & Yew, 2020) and studied changes in teachers’ assessment literacy resulting from professional development activities (Gotch & McLean, 2019). Research has also evaluated the technical adequacy of assessment literacy measures (DeLuca et al., 2016b; Gotch & French,

2014). These measures can provide teacher educators and staff from state and local education agencies with information they can use to improve areas of misunderstanding, particularly if applied to a specific measurement context (e.g., the state summative achievement assessment).

Despite the prevalence of assessment literacy materials and methods for evaluating their effectiveness, the materials may not adequately educate users about all assessment contexts. The availability of materials specific to certain assessments (e.g., classroom assessment; Evans & Thompson, 2020) suggests that assessment literacy needs may differ by context rather than being a single universally-defined competency. In general, materials typically cover traditional assessment concepts (e.g., differences between summative and formative assessments, meaning of a total scale score) and may be too narrow to apply to diagnostic assessment contexts. In the absence of materials specific to diagnostic assessments, educators risk applying traditional measurement concepts and using diagnostic assessment results based on incorrect assumptions.

Assessment Literacy for Diagnostic Assessments

Diagnostic assessments are known for providing fine-grained mastery results (Leighton & Gierl, 2007). Rather than report a raw or scale-score value that represents student performance on the full set of measured attributes or skills, diagnostic assessments report student performance relative to each skill measured by the assessment. Results summarize the likelihood that students have mastered skills or skill profiles, typically in the form of probabilistic values (e.g., calculated via diagnostic classification modeling; Bradshaw, 2017) or dichotomous mastery determinations based on a cut point (Karvonen et al., 2019).

While these detailed skill-level mastery profiles are more informative to instructional practice than traditional results (Clark et al., 2018), diagnostic assessments also depart from traditional measurement concepts that teachers are likely familiar with. Said another way, what

constitutes a “basic” assessment concept (Popham, 2011) may differ. For instance, terms like *score* and *subscore* are replaced with terms like *skill mastery* and *profile of skills mastered* in a diagnostic assessment context. Teachers may be unfamiliar with how skill mastery is determined or unsure how confident they can be that mastery determinations reflect students’ actual achievement. Classification terms for mastered and not-mastered skills may mislead teachers (Bradshaw & Levy, 2019), and visual representations of the certainty of mastery determinations can be misinterpreted (Feldberg & Bradshaw, 2019). Thus, diagnostic assessments may challenge the boundaries of traditional teacher assessment literacy (Leighton et al., 2010); however, there is limited research to date examining teachers’ knowledge for diagnostic systems.

Assessment Literacy: Interpretation and Use

Critical to assessment literacy is one’s ability to understand assessment concepts and their impact on instructional decision-making. Stiggins (1991) introduced the concept of assessment literacy to describe individuals’ ability to evaluate assessment quality and determine whether the format of results promotes interpretation and use. Popham (2011) further stipulated that assessment literacy pertains to the most critical assessment concepts that teachers need to know to interpret results. Assessment literacy is subsumed within the broader data literacy umbrella; that is, teachers’ ability to use data from a variety of sources to inform instruction, including assessment data (Mandinach & Gummer, 2016). Each of these definitions emphasizes that teachers’ interpretation and use of results is an important component of assessment literacy.

Improved assessment literacy has the potential to expand the ways educators use data to inform instruction and can support them in making better instructional decisions (Guskey, 2020). Using data to inform subsequent instruction is effective at improving student learning (Hamilton et al., 2009; Wiliam, 2011). Teachers indicate that access to assessment results affects their

instructional practice (Datnow et al., 2012), and research suggests that teachers' ability to make sense of data strongly influences their use of the information (Cho & Wayman, 2014). Therefore, assessment literacy, and specifically being able to understand and use assessment results, is a critical contributor to student learning and has direct connections to the intention of diagnostic assessments (i.e., fine-grained reporting supports instructional decision making).

Because assessment literacy requires teachers to interpret and use results, results must be presented in a usable format. Score reporting literature details the importance of reports being easy to interpret and use (Hambleton & Zenisky, 2013) and emphasizes the critical link between teacher understanding and use of results and assessment validity (American Educational Research Association [AERA] et al., 2014; Tannenbaum, 2019). Score reports are intended to provide educators with "the information they need, in a way that they understand, so that they may reasonably act on that information" (p. 11, Tannenbaum, 2019) and bridge the information the assessment measures and teachers' instructional decisions (Zapata-Rivera & Katz, 2014). That is, score reports should support assessment literacy. Further, because score reports communicate information about the assessment in addition to results, "it follows that score reports may be the gate keepers to test users' ability to understand and interpret data" (Ketterlin-Geller et al., 2018, p. 1) and can serve as a mechanism for discussing and understanding teachers' assessment literacy (Kim et al., 2020).

Purpose

This study explored teachers' diagnostic assessment literacy. We examined how teachers interpreted and used diagnostic score reports and their broader understanding of a diagnostic assessment system. We drew from Popham's assessment literacy definition when structuring research questions specific to understanding diagnostic assessment literacy, including:

1. How do teachers talk about fundamental concepts related to diagnostic assessments?
2. Do teachers demonstrate understanding of the diagnostic system and its results?
3. How does diagnostic mastery information inform instructional decision-making?

Method

We collected data from score report focus groups conducted with teachers to understand their assessment literacy for a diagnostic system and supplemented focus group data with teacher survey responses (e.g., Morgan, 2004). We used qualitative analytic methods to code focus group transcripts for content related to all three research questions and summarized survey item descriptive statistics to answer the second research question.

Study context

We answered the research questions in the context of the Dynamic Learning Maps[®] (DLM) Alternate Assessment System, which is an operational assessment system that delivers diagnostic alternate assessments used in 18 states (at the time of the study) as their large-scale academic assessments for students with significant cognitive disabilities. DLM assessments define extended content standards for each grade (3 through 8 and high school) and subject (English language arts, mathematics, and science). For each standard, students are assessed after instruction on one or more complexity levels. In English language arts and mathematics, there are five levels; in science there are three. Rather than a fixed form assessment, content is delivered in a series of short assessments, called testlets, each measuring one standard and level. Some parts of the assessment are computer-administered with teacher support as needed, and others are directly administered by the teacher outside the online system.

Student responses to DLM assessments are scored using diagnostic classification modeling (DLM Consortium, 2018). For each assessed level, the scoring model determines

dichotomous student mastery (i.e., mastered or not) based on the student's mastery probability. Unlike scale-score based assessments, which produce a total score that may be broken into subscores, the DLM system produces separate mastery determinations for 27 to 70 levels across the blueprint (depending on grade and subject) and aggregates up from mastery determinations for summative uses. Mastery information is summarized in individual student score reports, which were refined through a series of interviews and focus groups (Clark et al., 2015; Karvonen et al., 2016; Karvonen et al., 2017). Student reports include two parts: a Learning Profile (Figure 1) and a Performance Profile (Figure 2). The Learning Profile provides fine-grained skill mastery for each assessed content standard and level. The Performance Profile aggregates skill-mastery information across (a) sets of conceptually related standards, shown as the percentage of skills mastered per area; and (b) the subject overall, which uses a mastery-profile-based standard-setting method to set cuts between total levels mastered (Clark et al., 2017). Because of student population heterogeneity, teachers are accustomed to making individualized instructional decisions, so the Learning Profile is more immediately applicable for planning purposes.

Prior to test administration, teachers complete or annually update required training, covering a range of topics including assessment design, administration, and scoring. Teachers learn that the assessment is scored using a nontraditional scoring system that determines results based on the likelihood of skill mastery. The DLM website provides optional supplementary documents and videos to explain score-report contents and guide teachers when discussing assessment results and score reports with parents.

Data Collection

Data were collected from two sources: focus-group interviews and a teacher survey.

Focus Groups

Researchers developed a protocol in advance of the meetings. The protocol included questions intended to gauge participants' understanding of the diagnostic assessment system and the results produced, and how they used diagnostic results to inform instruction. Focus group questions covered four areas: 1) general background information, 2) preparing for and using the assessment system, 3) training and resources on the assessment and score-report contents, and 4) participants' interpretation and use of score reports for instructional decision-making.

Following IRB approval, we recruited teachers to participate in focus groups. We asked all 18 consortium state education agencies to distribute focus group information. Interested individuals provided their information via a Qualtrics survey (www.qualtrics.com). Because focus groups were structured around score report interpretation and use, to be eligible to participate, teachers had to have administered DLM assessments and received and used score reports for the prior academic year. We received responses from 170 interested individuals. Of those, 132 provided the required information, and 40 were eligible to participate. We emailed the eligible teachers and asked them to indicate their availability to participate in a session. Thirty teachers agreed to participate. We conducted eight focus groups, with 17 teachers, due to attrition challenges between scheduling and conducting the meetings. This resulted in the number of participants per event ranging from one to five; several focus groups were conducted as one-on-one interviews. They are collectively referred to as focus groups throughout.

The 17 participating teachers mostly self-reported as White ($n = 13$) and female ($n = 13$). Teachers taught across rural ($n = 2$), suburban ($n = 9$), and urban ($n = 5$) settings in three states. Teachers reported a range of experience with subjects and with students with significant cognitive disabilities. Most teachers taught more than one subject, and their collective experience spanned all tested grades. Teachers indicated they taught between one ($n = 3$) and 15 or more (n

= 2) students taking DLM assessments, with most ($n = 8$) reporting between two and five.

Prior to each focus group meeting, example score reports were shared with participants. Although all participants indicated they had received and used score reports, the templates were shared to provide a common example from which to frame discussion. Focus groups were conducted virtually using Zoom videoconferencing. At the beginning of each focus group meeting, the facilitator reviewed informed-consent information and advised that the session would be recorded. Focus groups followed a semi-structured format that included questions from the protocol. As additional topics emerged from the participant discussion, the facilitator incorporated probing and/or referencing score-report templates. Sessions lasted approximately 90 minutes, and each participant received \$50. Audio from each session was transcribed verbatim by an external professional transcriber for subsequent analysis and were reviewed for accuracy.

Teacher survey

A teacher survey is annually assigned to all teachers administering DLM assessments. Teachers access the survey via the online test administration portal between March and June. State and local education agencies encourage participation. The survey first advised teachers to administer assessments to students before completing the survey and presented an informed consent notice. Teachers indicated their willingness to participate by responding to the survey. Surveys took approximately five to 10 minutes to complete, and teachers could exit at any time.

Eight survey forms were available during the spring 2018 administration. Survey forms were randomly and equivalently assigned: teachers had an equal likelihood of receiving any given survey form. Teachers who had multiple students were assigned multiple surveys (i.e., one per student experience). One survey form included five items that asked teachers to rate their experience with training and resources, using the online system, and administering assessments

to the student. Items had 4-point Likert-scale response options (i.e., *strongly agree* to *strongly disagree*). Items were written by research team members and reviewed by consortium state education agencies, the Technical Advisory Committee, and IRB prior to administration.

A total of 19,144 teachers (78.0%) from all states responded to a survey form for 53,543 student experiences (62.3%). Of those, 8,416 teachers received and responded to the form with the teacher experience questions about 12,289 student administrations. These teachers represented 2,567 (67%) districts, and 5,855 (50%) schools. Teachers mostly reported having two or more years of experience with the system (76%), and the most often reported amount of teaching experience in any subject was 0-5 years (30-36%), which were both consistent with the distributions in the full teacher population.

Data analysis

We developed an initial set of codes and definitions as they related to each research question from our knowledge of the current research literature and listening to the focus groups. We applied the initial codes to short segments of text for one transcript and then met to discuss and reconcile any differences, clarifying definitions and adding codes as needed. We coded one more transcript and again met to discuss; no further codebook changes were made. The codebook had 18 codes across three categories, including assessment (codes included use of assessment data, interpretation of results, etc.), instructional practice (codes included planning, decision-making, etc.), and teacher (codes included self-efficacy, training, etc.). We independently coded the remaining transcripts and applied updated codes to the first transcript. We read each transcript multiple times to become immersed in the content. When coding, we first identified segments of text associated with each research question, then read and coded again, applying the content codes. Codes were associated with segments of text using Dedoose

qualitative data-analysis software, and code reports were generated for each research question to compile excerpts across focus groups. Researchers used the transcripts and code reports to identify code patterns (Creswell & Poth, 2018) and summarize findings by research question.

Data from the teacher survey were compiled following the close of the spring administration window. We used teacher-survey data to help answer the second research question about understanding the system, combining it with the qualitative evidence from the focus groups. Frequency distributions were provided for each survey item.

Credibility and trustworthiness were established using several methods. We included teacher perspectives from a variety of states, backgrounds, and classroom settings as a form of perspective triangulation (Patton, 1999). While conducting focus groups, we asked clarifying questions and paraphrased responses to verify our understanding. We relied on a team of researchers to code verbatim focus-group transcripts and used a codebook to promote consistency. We supplemented focus group feedback with teacher survey data.

Positionality

The research team included the three co-authors, all of whom hold doctorates in educational psychology and research and work on the DLM project. One researcher conducted all focus group sessions, and one observed all sessions. Two researchers prepared and refined the coding protocol and coded focus group transcripts, with support from an additional staff member who also holds a doctorate in educational psychology. A research associate with a master's in applied statistics retrieved and summarized the teacher survey results. A research assistant with a master's degree in creative writing combined transcription codes and retrieved code reports.

Results

Findings are summarized for each research question, specifically, how teachers talked

about diagnostic assessment concepts, their understanding of the assessment design and results, and their use of results to inform instruction.

Diagnostic Assessment Concepts

Focus group data revealed that the most salient concept related to diagnostic assessments is that of skill mastery. Teachers readily adopted and were comfortable using mastery terminology to talk about assessment results and student knowledge. Teachers described using reports to “see if they’ve *mastered* this” and to talk more generally about students’ learning (e.g., “even though they’re not *mastering* the same things at the same time, they’re very similar in where they are in their learning”). Teachers described the benefits of receiving fine-grained mastery information and how it contrasted with results from other assessments that provide coarser results. As one stated, the fine-grained mastery information gives “a more honest assessment of where students stand.” As another described, “That’s why I say it’s so much better than the old system because it tells me exactly what they’ve mastered and not mastered.” These comments reflect teachers’ conceptions of skill mastery and that diagnostic assessments provided a way of understanding relative strengths and weaknesses.

Teachers also generally expressed agreement with the mastery results provided, suggesting they were consistent with classroom observations of skills students could demonstrate. One teacher described discussing the mastery results with another special education teacher and the extent results reflected the students’ actual knowledge.

I told him, “I want you to look at those reports and see if you really feel like that that is a reflection of where that student is.” And he looked at it, and he said, “Absolutely. That’s absolutely amazing.” [The results] are right on target when it comes to the shaded [mastery] areas [on the report] on where the students truly are. I think it’s a very good match. With the portfolio [alternate assessment], you didn’t get near what we get with this.

Despite their ease of using mastery language to talk about results, teachers' comments did not reflect a nuanced understanding of what mastery represented. Teachers tended to talk about mastery results as absolute; they did not talk about mastery as reflecting skills that were *likely* mastered or as being based on probabilities of mastery. As one teacher described, she views that the mastery information indicates to her "okay, yeah, he's really got that. So now let's move on." Only one teacher mentioned reviewing previously mastered concepts to make sure skill mastery was retained. Because teachers tended to agree with the mastery results presented on score reports, it was unclear whether teachers understood that, like all assessment results, mastery determinations contain some amount of measurement error and could potentially report mastery for a skill the student has not truly mastered, and vice versa.

It was also evident from their comments that teachers were unsure how mastery was defined or determined. Some teachers described more traditional definitions of mastery when describing score report results. They gave examples of a student demonstrating the skill with at least 80% accuracy, or four out of five trials, as is common for evaluating instructional objectives for students with disabilities. Others referred to the mastery shading or summaries of mastered skills as what students "got right." Other teachers shared confusion about how results were calculated. Teachers stated that explaining the mastery determination on score reports is a "very complex" thing and that they did not understand what happens after assessment administration to produce the mastery decisions or overall performance level in the subject. One teacher referred to the scoring process as a "black box." Another wondered whether sometimes the overall results for her student were based on a "lucky guess" the student made, which reflects a level of misunderstanding about how scoring works, implying that a lucky guess is enough for the scoring model to reflect broad skill mastery. These statements reflect a lack of familiarity

with and understanding of the diagnostic scoring model, which may appear more complex. However, it is also unclear if these more nuanced understandings about how mastery is calculated constitute fundamental understandings that are critical for appropriate use of results.

Understanding of Assessment and Score Report Contents

Focus-group participants described the diagnostic assessments in ways that reflected a general understanding of the assessment design and administration. They described being comfortable administering assessments in the online system and familiar with the content standards and assessment targets (i.e., what the assessment measured and at what level of complexity). Similar findings were observed in the teacher-survey responses. Teachers reported that they understood how to use the system and felt prepared to administer assessments. Teachers agreed or strongly agreed that they were confident administering assessments (97.0%), the required test-administrator training prepared them for administration responsibilities (91.2%), the manuals and resources helped them understand how to use the system (91.0%), they knew how to use accessibility supports and options for flexibility (94.5%), and the brief guidance documents accompanying each assessment helped them with delivery (90.1%). These findings suggest that teachers do not view the diagnostic assessment as being challenging to implement despite differences from traditional fixed form measures.

When discussing the score reports, teachers' statements generally reflected comprehension of their contents. Participants described using the Learning Profile to understand the skills students demonstrated. They correctly understood and described the colored shading on the report, which indicated for each content standard the levels mastered, not mastered, or not assessed, and they incorporated those interpretations into their discussion of the results. As one teacher stated, "I might have a kid that would be just blue [i.e., no evidence of mastery] instead

of the green [i.e., mastered], so I might still be addressing that particular [standard].” When describing the aggregated information on the Performance Profile, teachers correctly indicated it provided overall summary information for the subject, although many also made statements about the student’s overall “score” when referring to the performance level or their “score” on specific standards, despite the absence of quantitative values. Some teachers misinterpreted bar graphs on the Performance Profile that summarize the percentage of skills mastered for related content standards, incorrectly describing the results as the percentage of items a student correctly responded to or the percentage of trials in which the student demonstrated a behavior. These statements may suggest a reliance on more traditional assessment literacy concepts, such as receiving an overall score and results being based on a percent correct value; or connecting their interpretation to instructional practice, which often includes data collection based on percent of trials on which students demonstrated a correct response.

Despite understanding the assessment design and administration and having a general understanding of score report contents, all teachers participating in the focus group indicated a desire for additional training and resources, and specifically materials focused on understanding assessment results, which is a critical component of assessment literacy. Teachers who received local training indicated that it often prioritized assessment administration (i.e., areas they were comfortable with) and did not provide information on understanding score-report contents. One teacher described her first year using the assessment system as “drinking from a firehose,” referring to the flood of information the assessment training provided.

That first year, in a firehose scenario, [the score report] wasn’t very meaningful. I didn’t get a lot out of it. I wasn’t able to give the parents a lot out of it other than, “Here’s your score report. It’s color-coded so you can see where your kid [mastered skills].”

This statement likely reflects more rudimentary assessment literacy for a diagnostic system.

However, the teacher went on to say that she knew she could get more out of the score report and independently visited the assessment website to find additional resources. Only then did she feel confident discussing results with parents in a way that reflected a deep understanding of the assessment system, the results, and how both the system and results connected to her instruction on the content standards and a student's IEP goals. These comments are indicative that more nuanced assessment literacy developed over time but was the result of self-directed inquiry.

Instructional Use

To demonstrate assessment literacy for diagnostic systems, teachers should understand how results can be used to inform instruction. Regardless of teachers' differing conceptions of skill mastery and how it was calculated, the results provided in the diagnostic score reports certainly influenced teachers' instructional decision-making. Because teachers generally accepted the mastery determinations as true, they appeared to place a high degree of trust in the results when planning next steps for instruction.

Teachers reported that skill mastery information was beneficial for creating subsequent instructional plans that aligned to students' next steps for learning. Several teachers referred to using the fine-grained mastery results to identify gaps in students' skills that would be the focus of subsequent instruction, specifically using the information in the Learning Profile to determine which levels had been mastered for each content standard and which were still needed to reach the grade-level target. One teacher summarized her approach to using results as, "if the student didn't really show mastery, here's some information that I can use to understand how to go about thinking about instruction next." Another teacher said she could use the mastery results "to re-introduce those skills to make sure they're not falling behind." Teachers varied in prioritizing greater instructional depth in a particular area of related standards or wider breadth across areas

based on students' performance, both of which are appropriate and intended uses of results.

In addition to planning student-specific instruction, teachers described examining patterns of performance across students to improve their pedagogy. One high school teacher described:

What I do is I look at, in the past, I look at the score reports and said where were my students lacking? And if it was consistent like from one year to the next then I knew I had to enhance whatever I was doing in that component that they didn't do well [in]... I kind of [consider], well how did I teach it that year, could I have pushed something a little harder for these students?

This level of self-efficacy and personal accountability requires assessment literacy to be able to use results in the aggregate to identify areas for professional growth.

Teachers also reported that the mastery results were useful for planning student IEP goals. One said, "it's interesting because their IEP goals are very similar to what is their level [for the standard]. I can do that, I can say, 'Hey, let's look at this level and let's look at this Target skill [i.e, grade-level expectation] and this is what we're working on in your IEP.' It's real easy for me to tie all of these things together." Another teacher described using results at IEP meetings, saying, "I pointed out some of the IEP objectives and how they were related to what was on the report." Teachers recognized that the fine-grained diagnostic mastery information provided by the assessment supported them in creating goals individually suited to the student while also reflecting high expectations, as intended by the diagnostic system.

However, teachers did also express a desire for additional information about how to use the results for instructional decision-making. One teacher described the lack of information shared with her by the local school district:

I think [the score report] was given to us as something to give to parents. I don't think it was given to us as a, "You should glean anything out of it. You should be able to use this to make good educated choices on the direction you take a kid..." Somebody at some point should sit

down with all the teachers who gave [the assessment], with our score reports, and say, “Hey let’s look this over. Let’s see what we can get out of this.”

Because mastery information differs from results obtained from traditional scaled score and subscore reporting, teachers may desire additional support and resources to feel confident they are using results correctly.

Discussion

Research into assessment literacy has largely focused on traditional assessment contexts. As the use of assessments scored with diagnostic models expands from research applications to high-stakes purposes, it is important for the measurement community to explore the extent to which teachers understand diagnostic assessment results and how they can be used to plan instruction, due to differences in the reporting structure. This study bridges theory and practice by examining assessment literacy, and specifically how teachers talk about diagnostic assessments, their understanding of the assessment and results, and how results inform instructional planning, for one of the few diagnostic assessment systems in operational use.

Interpreting Mastery and Implications for Reporting

The concept of reporting mastery results is unique to diagnostic assessments and requires teachers to have different assessment-specific knowledge to interpret and use results. This includes identifying strengths and weaknesses relative to targets and being able to differentiate that from interpretation of an overall achievement indicator used for making normative comparisons. Participants in our study easily adopted the mastery language of diagnostic assessments when interpreting and using results. This may be due in part to the study context. Special educators administering DLM assessments may already be accustomed to using mastery language to describe performance, for instance when evaluating student achievement of IEP goals, despite alternate assessment results largely prioritizing deficits over strengths. We believe

the structure of diagnostic score reports and use of mastery shading support teachers in adopting mastery-based language to talk about assessment results. This is consistent with other studies examining teachers' use of diagnostic score reports that suggest they are easy to interpret and use (Feldberg & Bradshaw, 2019; Karvonen et al., 2017).

Because mastery is such a common term that teachers have some familiarity with, prior conceptions could potentially influence their interpretation of results. This is contrasted with traditional assessments that report results as raw or scaled score numeric values; while teachers may similarly not understand how a scaled score is calculated, because the terminology differs from words used to describe student achievement in their classrooms, it may be accepted without the same confounding of terms. The extent that the operational definition of mastery differs from teachers' conceptions of how students demonstrate mastery (e.g., percentage of correct item responses, percentage of trials) may impact instructional decisions. For instance, if a teacher believes mastery is determined by a student correctly answering all items measuring the skill, they may ascribe greater certainty to the results than a probabilistic mastery determination intends. However, having a general sense of what mastery constitutes, even if the teacher is not exactly sure how it is determined, may constitute *enough* assessment literacy to be able to use results. In other words, it may not be of consequence if teachers believe that mastery is based on a percentage correct versus a probability, so long as they can use mastery information to plan next instruction. Future research could explore how much assessment literacy, and what specific information, is sufficient for stakeholder use of results from a diagnostic assessment and whether reports should include additional information (e.g., what constitutes mastery).

Probing during focus groups also revealed a lack of teacher understanding of how dichotomous mastery, represented on the reports as mastery shading, was determined. While a

lack of nuanced understanding of assessment results is not unique to diagnostic assessments, this finding does have implications for the design of diagnostic assessment score reports. Test developers must decide whether to report probability values or dichotomous mastery statuses. Early focus-group feedback on score-report prototypes (Clark et al., 2015) indicated that teachers preferred mastery-status designations over probability values because they are easier to interpret. Other studies have found that teachers, even after training, may incorrectly interpret the probability value as a percentage correct (Feldberg & Bradshaw, 2019). Recognizing that teachers prefer mastery statuses and that they are easier to interpret, but that teachers tend to treat them as absolute rather than a likelihood, test developers can support teachers in making statistically supported decisions. Examples include setting mastery thresholds further from the point of maximum uncertainty (i.e., .5, which is a common threshold in the literature) and closer to maximum certainty (i.e., 1.0), and by including interpretive information on score reports about what uncertainty means in this context. This internally addresses (i.e., within scoring and reporting) an area that might pose challenges to teachers' assessment literacy rather than requiring or assuming teachers understand this distinction.

Teacher conceptions about the diagnostic scoring model may also impact the instructional decisions made from mastery results. For instance, one teacher's expression of the results perhaps reflecting a student's "lucky guess" likely demonstrates a lack of understanding about the relationship between scoring and the number of items measuring a skill, and may also reflect low expectations for her own students' capability to demonstrate academic skills (which is a historic challenge for students who take alternate assessments; e.g., Timberlake, 2014). Teachers are not likely to know that the assessment design and scoring model account for guessing (non-masters provide correct item responses) and slipping (masters provide incorrect

item responses). Teachers who believe students can make a lucky guess on an item and be deemed a master of the skill may be distrustful of the results and be less likely to use them. Test developers should design diagnostic assessments to have sufficient items measuring each skill such that a single lucky guess would not result in incorrect mastery status or a single mistake result in non-mastery status. Developers should also consider whether score reports can make this information evident to support teachers in making appropriate interpretations from results.

Finally, teachers' use of *score* to describe student performance on diagnostic assessments despite the absence of a numerical value describing performance may also reflect some reliance on their broader assessment literacy and traditional conceptions of reporting. It requires a paradigm shift for teachers to begin thinking of results as a comprehensive profile of mastered multidimensional skills rather than a single descriptive number or value of a unidimensional academic construct. However, as indicated by the focus-group responses, fine-grained mastery profiles support connections to instructional practice beyond results typically delivered by traditional large-scale assessments. Because skill mastery is reported in relation to the target and achievement on other standards, the reporting format likely supports teachers in planning subsequent instruction even if they have less content knowledge or pedagogical content knowledge (e.g., special educators; Brownell et al., 2017). This reporting structure also departs from item-level information sometimes provided by traditional large-scale assessments, which require educators to translate item-level responses up to conceptual thinking about the assessment target. In these ways, reports from diagnostic assessments can support educators and can be valuable tools in and of themselves in promoting teachers' assessment literacy.

Resources for Supporting Assessment Literacy

Teacher conceptions along with explicit requests for additional training emphasize a need

to make information about diagnostic assessments clear and readily available. While the DLM Consortium makes several resources available to teachers to support understanding of how mastery is determined through required training, helplet videos, and FAQ documents, further exploration may be needed to bridge the gap between the contents of available resources, including score reports themselves, and teachers' current level of understanding. Additional materials could be made available to district staff to support professional development and professional learning communities to help broaden the current teacher-described focus on administration to also include reporting and use of results. By providing materials targeted at common areas of confusion and making them easily accessible, the information may support teachers in understanding what results mean and how they can be used to inform instruction.

However, it is also important for test developers and education agencies to consider what constitutes sufficient assessment literacy to support adequate interpretation and use of results. With districts already busy and resources stretched thin, providing further teacher materials may not always be practical. Similarly, there needs to be a balance between meeting teachers' desires for more resources and recognizing that many have limited time to access them. Rather than placing the onus on teachers to recognize their need for additional materials, locate resources, and devote time to learning and using them, test developers can support teachers' assessment literacy in other ways. For instance, test developers can consider how much information to include on score reports versus in supplemental materials. They can conduct studies to determine the types of materials teachers find most useful and consider the ease of locating materials and how to best communicate about their availability to stakeholders.

Fostering assessment literacy for diagnostic assessments appears to be a balance between supporting educators inherently within the system and score reports, while also building their

knowledge of the truly fundamental concepts related to diagnostic assessments. Additional research may help identify what are, in fact, fundamental concepts versus supplemental or nice-to-have knowledge (i.e., whether there are tiers of assessment literacy for diagnostic assessments). For instance, additional studies could shed light on how critical it is for teachers to understand how probability-based mastery statuses were determined or uncertainty in mastery determinations to be able to use them to effectively inform instruction.

Additional Considerations

In addition to considering teachers' assessment literacy, the adoption of diagnostic assessment systems may require building assessment literacy in other stakeholder groups, particularly because of the unique ways in which diagnostic assessments differ from traditional scale score measures. For instance, when legislators and policymakers do not adequately understand an assessment, they may make ill-informed decisions regarding its administration and use (Guskey, 2020). Similarly, parents may be less familiar with diagnostic mastery-based reporting. Providing adequate information to all relevant stakeholders will support proper interpretation and use of results at each level results for which results are intended to be used.

Limitations

Results of this study are based on self-report data primarily collected from teacher focus groups with limited sample sizes, constrained to teachers who reported receiving and using score reports. While teacher-survey results were included to supplement the findings gleaned from the focus groups for the second research question regarding teachers' understanding of the system, qualitative data were favored to provide a rich understanding of how teachers talked about diagnostic assessments and use of results. We recognize that a different sample may provide different responses that may lead to different conclusions. Similarly, teacher self-report via the

focus groups and surveys may have been biased. Although teachers appeared to be comfortable disclosing areas of less understanding, their responses may not have been fully representative. Alternately, the teachers we spoke to may have been more interested in use of score reports, given they indicated they had used them, and may have had more understanding than a typical teacher. However, we believe the findings we obtained advance the limited literature available on teachers' assessment literacy for diagnostic assessments and serve to inform future practice.

Implications for Practice

Assessment literacy, score reporting, and validity literature all emphasize the criticality of correct interpretation and use of results. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) encourage test developers to consider and promote assessment literacy and advise the collection of consequential evidence of the extent to which results are used as intended. Score report design should consider many factors, including balancing the types of information different stakeholders desire (e.g., accuracy versus clarity of information; Malone, 2013), that can also promote end users' assessment literacy. As part of these efforts, it is the responsibility of test developers to determine whether score report design and evidence of teachers' assessment literacy support appropriate interpretation and use of results.

Assessments scored with diagnostic models are beginning to transition from primarily research applications to operational use for accountability purposes. To date there is limited research into teachers' assessment literacy for diagnostic assessments or design of score reports to promote assessment literacy. This study advances and connects assessment literacy and score reporting literature and offers actionable information to inform future development of assessment literacy materials to support interpretation and use of diagnostic assessment results, as well as design of the reports to inherently promote teachers' understanding and ability to use them.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: National Council on Measurement in Education.
- Bradshaw, L. (2017). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297–327). Malden, MA: Wiley.
- Bradshaw, L., & Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educational Measurement: Issues and Practice*, 38, 79–88.
- Cho, V., & Wayman, J. C. (2014). Districts' efforts for data use and computer data systems: The role of sensemaking in system use and implementation. *Teachers College Record*, 116(2).
- Clark, A. K., Karvonen, M., Kingston, N., Anderson, G., & Wells-Moreaux, S. (2015, April). *Designing alternate assessment score reports that maximize instructional impact* [Paper Presentation]. National Council on Measurement in Education, Chicago, IL.
- Clark, A. K., Karvonen, M., Swinburne Romine, R., & Kingston, N. M. (2018). *Teacher use of score reports for instructional decision-making: Preliminary findings* [Paper Presentation]. National Council on Measurement in Education, New York.
- Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed mastery profile method for setting standards for diagnostic assessment systems. *Educational Measurement: Issues and Practice*, 36(4), 5–15. <http://doi.org/10.1111/emip.12162>
- Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design*. SAGE.
- Datnow, A., Park, V., & Kennedy-Lewis, B. (2012). High school teachers' use of data to inform instruction. *Journal of Education for Students Placed at Risk*, 17, 247–265.
- DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016a). Approaches to classroom assessment

- inventory: A new instrument to support teacher assessment literacy. *Educational Assessment*, 21, 248–266. <http://doi.org/10.1080/10627197.2016.1236677>
- DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016b). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28, 251–272. <http://doi.org/10.1007/s11092-015-9233-6>
- Dynamic Learning Maps Consortium. (2018). *Technical manual update—integrated model*. Lawrence: University of Kansas, Accessible Teaching, Learning, and Assessment Systems.
- Evans, C. M., & Thompson, J. (2020). Classroom assessment learning modules. Dover, NH: National Center for the Improvement of Educational Assessment.
- Feldberg, Z., & Bradshaw, L. (2019). *Reporting results from diagnostic classification models to teachers* [Paper Presentation]. National Council on Measurement in Education, Toronto.
- Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, 33, 14–18.
- Gotch, C. M., & McLean, C. (2019). Teacher outcomes from a statewide initiative to build assessment literacy. *Studies in Educational Evaluation*, 62, 30–36.
- Guskey, T. R. (2020). The dark side of assessment literacy: Avoiding the perils of accountability. *AASA Journal of Scholarship & Practice*, 17(1), 7–15.
- Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association.
- Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., Wayman, J., Pickens, C., Martin, E., & Steele, J. L. (2009). *Using student achievement data to support instructional decision making* (IES Practice Guide; NCEE 2009-4067). U.S. Department of Education.
- Illinois State Board of Education. (2015). *Guiding principles for classroom assessment*. <https://www.isbe.net/Documents/guiding-principles.pdf>

- Kansas Department of Education. (2019). *Assessment literacy project*. <https://www.k-state.edu/ksde/alp/>
- Karvonen, M., Clark, A. K., & Kingston, N. (2016). *Alternate assessment score report interpretation and use: Implications for instructional planning* [Paper Presentation]. National Council on Measurement in Education, Washington, D.C.
- Karvonen, M., Clark, A. K., Swinburne Romine, R., & Kingston, N. (2019). *Development and evaluation of diagnostic score reports for an alternate assessment system* [Paper Presentation]. American Educational Research Association, Toronto.
- Karvonen, M., Swinburne Romine, R., Clark, A. K., Brussow, J., & Kingston, N. (2017). *Promoting accurate score report interpretation and use for instructional planning* [Paper Presentation]. National Council on Measurement in Education, San Antonio, TX.
- Ketterlin-Geller, L., Perry, L., Adams, B., & Sparks, A. (2018). *Investigating score reports for universal screeners: Do they facilitate intended uses?* [Paper Presentation]. National Council on Measurement in Education Classroom Assessment Conference, Lawrence, KS.
- Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: A focus on educators of K–12 English learners. *Language Testing*, 37(1), 54–75. <http://doi.org/10.1177/0265532219859881>
- Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessments for education: Theory and applications*. New York, NY: Cambridge University Press.
- Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy, & Practice*, 17, 7–21.
- Lian, L. H., & Yew, W. T. (2020). Development of an assessment literacy super-item test for assessing preservice teachers' assessment literacy. *International Journal of Innovation, Creativity, and Change*, 13(7), 870–889.
- Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, 30(3), 329–344. <https://doi.org/10.1177/0265532213480129>
- Mandinach, E., & Gummer, E. (2016). What does it mean for teachers to be data literate: Laying


- out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, 60, 366–376.
- Michigan Assessment Consortium. (2020). *Assessment learning modules*. <https://www.michiganassessmentconsortium.org/almodules/>
- Morgan, D. (2004). Focus groups. In S. N. Hesse-Biber & P. Leavy (Eds.), *Approaches to qualitative research: A reader on theory and practice* (pp. 263–285). Oxford University Press.
- Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5 Pt 2), 1189–1208.
- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator’s confession. *The Teacher Educator*, 46(4), 265–273. <http://doi.org/10.1080/08878730.2011.605048>
- Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534–539.
- Tannenbaum, R.J. (2019), Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting Research and Applications* (pp. 9-18). Routledge, NY
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30, 251–275.
- Timberlake, M. T. (2014). Weighing costs and benefits: Teacher interpretation and implementation of access to the general education curriculum. *Research and Practice for Persons with Severe Disabilities*, 39, 83–99. <https://doi.org/10.1177/1540796914544547>
- United States Department of Education. (2015). *Assessment design toolkit*. <https://www2.ed.gov/teachers/assess/resources/toolkit/index.html>
- U.S. Department of Education. (2020). Innovative Assessment Demonstration Authority (IADA). <https://www2.ed.gov/admins/lead/account/iada/index.html>
- William, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 3–14.
- Wisconsin Department of Public Instruction. (2019). *Assessment literacy module*. <https://dpi.wi.gov/strategic-assessment/professional-learning/>
- Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, 21(4), 442–463. <https://doi.org/10.1080/0969594X.2014.936357>

Figure 1. Learning Profile summarizing skill mastery.

REPORT DATE: 06-06-2018
SUBJECT: English language arts
GRADE: 10

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

Individual Student Year-End Report
Learning Profile 2017-18



DISTRICT ID: DLM District
STATE: DLM State

Student's performance in 10th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during the 2017-18 school year. Grade 10 had 19 Essential Elements in 4 Conceptual Areas available for instruction during the 2017-18 school year. The minimum required number of Essential Elements for testing in 10th grade was 10. Student was tested on 17 Essential Elements in 4 of the 4 Conceptual Areas.

In order to master an Essential Element, a student must master a series of skills leading up to the specific skill identified in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

Area	Essential Element	Level Mastery				
		1	2	3	4 (Target)	5
ELA.C1.2	ELA.L.9-10.4.a	Identify familiar objects through property word descriptors	Identify definition of words	Identify missing words using sentence context	Use semantic clues to identify word meaning	Use semantic clues to identify phrase meaning
ELA.C1.2	ELA.L.9-10.5.b	Draw conclusions from category knowledge	Identify the multiple meanings of a word	Identify word meaning of multiple meaning words using context clues	Identify the intended meaning of multiple meaning words	Understand how multiple meaning words can result in humor
ELA.C1.2	ELA.RI.9-10.1	Identify concrete details in a familiar informational text	Identify concrete details in an informational text	Cite textual evidence for inferred information	Discriminate between citations for explicit and inferred information	Cite evidence for a text's specific meaning

Levels mastered this year
 No evidence of mastery on this Essential Element
 Essential Element not tested

Page 1 of 4

©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.

Figure 2. Performance Profile summarizing overall achievement.

