Condensed Mastery Profile Method for Setting Standards for Diagnostic Assessment Systems

Amy K. Clark*

Brooke Nash

Meagan Karvonen

Neal Kingston

Center for Educational Testing and Evaluation

University of Kansas

The present publication was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the authors, and no official endorsement by the U.S. Department of Education should be inferred.

*Corresponding author

The Final Published Version of this manuscript is available in Educational Measurement: Issues and Practice, July 12, 2017 https://doi.org/10.1111/emip.12162

Abstract

The purpose of this study was to develop a standard setting method appropriate for use with a diagnostic assessment that produces profiles of student mastery rather than a single raw or scale score value. The Condensed Mastery Profile Method draws from established holistic standard setting methods to use rounds of range finding and pinpointing to specify cut points between performance levels. Panelists are convened to review profiles of mastery and specify cut points between performance levels based on the total number of skills mastered. Following panelist specification of cut points, a statistical method is implemented to smooth cut points over grades to decrease between-grade variability. Procedural evidence, including convergence plots, standard errors of pinpointing ratings, and panelist feedback, suggest the Condensed Mastery Profile Method is a useful and technically sound approach for setting performance standards for diagnostic assessment systems.

Keywords: standard setting, diagnostic classification modeling, student profiles, body of work

Condensed Mastery Profile Method for Setting Standards for Diagnostic Assessment Systems

Historically, claims of diagnostic assessment scores have rested on use of sub-scores derived using traditional classical or item response theory approaches applied to a small number of items. More recently, the level of attention for diagnostic classification modeling—a family of approaches designed to focus on diagnostic information rather than a secondary analysis—has increased (e.g., Gierl & Cui, 2008; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010; Sinharay & Almond, 2007; Templin & Bradshaw, 2014). The draw of diagnostic classification modeling is in large part due to its ability to provide rich reporting of student performance (Huff & Goodman, 2007). Rather than yielding a single score value that characterizes their overall performance, diagnostic assessments provide profiles that include fine-grained information about the skills students have mastered. Educators and parents can use such detailed reports as the basis for instructional decision-making and to determine next steps for enrichment or remediation.

While diagnostic assessments provide rich information about the specific things students know and can do, their use has been largely restricted to small-scale research applications (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Broaddus, 2012; Skaggs, Hein, & Wilkins, 2016). Recently, a large-scale assessment system based on diagnostic modeling, the Dynamic Learning Maps (DLM®) Alternate Assessment (Kingston, Karvonen, Bechard, & Erickson, 2016), was launched. As diagnostic assessments transition from research-based applications to assessment systems implemented statewide and used for accountability purposes, standard setting methods must be developed to categorize such nuanced profiles of student learning into performance labels that can be used in state accountability metrics. This paper describes a method for setting standards for diagnostic assessments, using the DLM assessment system as an example.

The Condensed Mastery Profile Method is suitable for diagnostic assessment systems because it does not require using a raw or scale score value, or an item-based approach. The sections that follow detail how established standard setting methods were adapted to accommodate assessment results that are based on a profile of mastery statuses in order to determine performance-level classifications to be used in state accountability systems.

Background

Diagnostic Assessment

The construction of a diagnostic assessment system begins with the specification of the skills or attributes to be measured by the test (e.g., Leighton & Gierl, 2007). These attributes represent knowledge, skills, and abilities students can acquire over time. These skills are then ordered into hierarchies, progressions, or learning map models by a series of directional pathways that indicate the hypothesized order of skill acquisition. During the test development process, items are written to measure the attributes, or nodes, in the map model. Items are associated with the nodes by the specification of a Q-matrix, which is a table of dichotomous values associating each item with the nodes or attributes it measures.

The output of the diagnostic scoring method is the complete set of student mastery probabilities for each measured attribute. Student mastery probabilities for each attribute are determined by Expected a Posteriori (EAP) estimates. These EAP estimates represent the probability that a student has mastered each individual skill, where values closer to 0 or 1 represent greater confidence that the student has either not mastered or mastered the skill, respectively. Values near 0.5 represent maximum uncertainty in the student's mastery status. Mastery status can be further defined by specifying a mastery threshold, beyond which a student is considered a master of the skill. Threshold values can be specified by expert judgment or statistical analyses (Rupp, Templin, & Henson, 2010). For operational assessments, the specification of this threshold may be largely a policy decision whereby stakeholders must balance the desire for a high level of classification certainty with a value that is also attainable by students. To inform this decision, data can be provided demonstrating the percent of students who would demonstrate mastery based on varying thresholds (e.g., 0.5, 0.6, 0.7, 0.8, 0.9) and expert judgment can inform the final selection of a value based on knowledge of the student population.

Once a threshold value has been specified, the probability values can be translated to dichotomous mastery statuses. As such, the basis of reporting for diagnostic assessments is not a single total score, scaled score, or set of sub-score values, but rather the set of posterior probability estimates or the dichotomous mastery statuses for all skills being measured, which provides fine-grained and detailed information regarding what the student knows and can do. For more information on diagnostic classification modeling, see Rupp, Templin, and Henson (2010).

In many research applications, diagnostic classification models (DCMs) have been retrofitted to previously existing tests (e.g., Skaggs et al., 2016; Svetina, Gorin, & Tatsuoka, 2011; Wang & Gierl, 2011) rather than creating a diagnostic assessment system built specifically for the purpose of diagnosing student mastery of attributes. In instances where the model is retrofitted, traditional standard setting methods can be applied when specifying performance standards. However, diagnostic assessment systems that report student performance in the form of a mastery profile cannot use traditional standard setting methods. Conventional raw or scale scores are not available to use as the basis of a standard setting procedure. Similarly, because probability estimates are obtained for the skill being measured, item-based methods are not appropriate either. To address these challenges, a method for setting standards must be developed that is appropriate for assessment systems based on DCMs. We describe such a method in this paper, including a comparison to other methods, a summary of the procedure used to prepare mastery profiles, specify cut points, and evaluate the identified cut points.

Dynamic Learning Maps Assessments

The DLM Alternate Assessment System is a consortium-based program that delivers assessments to approximately 90,000 students with significant cognitive disabilities in fifteen partner states and two Bureau of Indian Education tribal schools. Assessments are available for English language arts (ELA) and mathematics and follow two separate blueprint testing models from which states can choose: a spring summative model with a standardized blueprint, called the year-end model; and a through-year model with a flexible blueprint, called the integrated model. Students in year-end model states are assessed on 4-7 testlets in the spring, whereas students in the integrated model have a flexible blueprint that encourages teacher choice on the number and level(s) of testlets that are administered throughout the year. Data from the two models are calibrated together prior to scoring the assessments.

Testlets of 3-8 items are available for every Essential Element (EE), or content standard, at one of five linkage levels, or levels of complexity. The Target linkage level represents the grade-level standard, with the other linkage levels providing variation from the grade-level target in breadth, depth, and complexity. There are three Precursor linkage levels leading up to the Target: Initial Precursor, Distal Precursor, and Proximal Precursor. In addition, there is one linkage level beyond the Target: Successor. Each linkage level represents one or more nodes, or skills, in the learning map model that underlies the assessment system. Figure 1 provides an example of the five linkage levels that are available for assessment for each EE and the one or more nodes measured in each linkage level.

[Insert Figure 1 about here]

The scoring model for DLM assessments is a DCM that makes use of latent class analysis to provide posterior probabilities of mastery for each linkage level. The scoring model assumes items are fungible within a linkage level, meaning that item parameters for the linkage level are the same for all items. Said another way, the Q-matrix for the linkage level contains a column of 1s. Calibration combines data from the two blueprint testing models to obtain item and structural parameters used for scoring. Model calibration and scoring are both done using a program developed in the R Project for Statistical Computing (R Core Team, 2013). Additional information on the specification and estimation of the model can be found in Dynamic Learning Maps Consortium (2016).

Reporting for DLM assessments is based on the student's mastery of linkage levels. Each linkage level is represented as a dichotomous mastery status, either mastered or not mastered, based on a threshold for mastery probability adopted by the consortium (0.8). This threshold was based on a combination of expert judgment and review of student data. More information on the process for specifying the threshold can be found in Karvonen, Clark, and Nash (2015a and 2015b). Because of the ordering of the linkage levels, students who have mastered a higher linkage level (e.g., Target) are assumed to have mastered all prerequisite linkage levels as well. For each assessment, there is a total possible number of linkage levels that can be mastered, which is determined by multiplying the number of EEs, or content standards, by the five linkage levels available for each. The total number of linkage levels mastered can then be used as the basis for setting performance standards. However, this value does not represent a traditional scale or raw score, since linkage levels across all the content standards vary in terms of both grain size and amount of skill acquisition needed to move from mastery at a lower linkage level to the next. To guard against misinterpretation, the term "results" rather than "score" is used in the context of DLM assessments, and the term "cut point" rather than "cut score" is used throughout to refer to the value that distinguishes two performance levels.

Standard Setting Methods

While there are many standard setting procedures that have been implemented across different testing programs (e.g., Cizek, 2012), the overarching goal is the same. The standard setting process is conducted to specify distinctions, or cuts, between categories that describe student performance. Examinees who perform above the cut are given one performance classification, and examinees below the cut are categorized to another. The number of cuts specified during standard setting varies based on the purpose of the assessment, and can range from one to many.

Many state education agencies rely on the results of standard setting and the associated performance categories to feed into their statewide accountability models. Students' performance classifications can also impact programmatic decisions at the state and local level. Because cut scores derived from standard setting determine the boundaries for the performance classifications, a minor difference in the results of the standard setting process can have serious repercussions for students and teachers.

The wealth of research available on standard setting methods has expanded since Nedelsky (1954) wrote about methods for absolute grading standards, including an expansion of resources to support implementation of best practice (e.g., Cizek, 2012; Hambleton & Pitoniak, 2006; Zieky, Perie, & Livingston, 2008). The latest version of the *Standards for Educational and Psychological Testing* also includes guidance on best practice, including standards specifying the need for clear documentation of methods, approaches that allow panel participants to make use of their knowledge and experiences, and the use of sound empirical data to inform the process (American Educational Research Association [AERA], American Psychological Association, and National Council on Measurement in Education, 2014). Because of the consequences associated with the standard setting outcomes, the standard setting method selected should, above all else, be appropriate for the assessment for which standards are being set.

The most common standard setting approaches in large-scale educational assessment (e.g., modified Angoff and Bookmark methods) rely on evaluation of test items to specify cut scores along a scale score continuum. In these methods, items are often ordered by their difficulty, and performance levels are set based on panelist judgments about the response demands of the items, given their knowledge of the content and the population of test takers. There are also many methods based on the categorization of student work samples into performance-level classifications (e.g., body of work, performance profile). These methods rely directly on student evidence and are appropriate when the assessment features a collection of evidence such as constructed response items or portfolios. Because diagnostic assessments provide a detailed profile of mastery that summarizes performance on a number of skills, holistic methods were examined as a starting point for selecting a method that would be appropriate for the assessment.

While there are a number of holistic standard setting methods, they have many commonalities between them. Well-qualified panelists are recruited and selected for panel participation and undergo training prior to setting standards for the assessment. In addition, many holistic standard setting methods make use of an iterative process, cycling through rounds of range finding, which involves identifying the general range in which the cut lies, and pinpointing, which determines the specific cut between performance levels.

Body of Work Method. The Body of Work Method (Kingston, Kahl, Sweeney, & Bay, 2001) relies on a complete set of student work to be presented to the standard setting panelists. This typically consists of student work samples for constructed response tasks and may also include multiple-choice items. Panelists consider the complete body of work when classifying the student to one of two or more performance categories. The Body of Work Method has been widely used for performance assessments and has been referenced as one of the most widely implemented holistic approaches to setting standards (Cizek & Bunch, 2007).

The Body of Work standard setting process involves a series of steps, typically including rounds of range finding followed by pinpointing, based on student work organized into folders based on total score. The complete Body of Work Method and supporting research is described in Kingston and Tiemann (2012). The method has typically been applied to performance- and portfolio-based assessments where student work is produced, with categorizations of work such as a writing product or other collection of work samples.

Generalized Holistic Method. The generalized holistic method, first introduced by Cizek and Bunch (2007), draws upon other standard setting methods, such as Body of Work, Contrasting Groups, Bookmark, and Analytical Judgment methods to set performance standards. The collection of evidence serving as the basis of the approach typically consists of multiple-choice items or student work samples. Similar to Body of Work, the collection of evidence is presented to panelists in order based on total score. Panelists use the analytical judgment procedure to evaluate student work and classify collections into the performance-level categories. Panelists then work through two rounds of range finding, eliminating the pinpointing stage from the standard setting process altogether.

Performance Profile Method. The Performance Profile Method is a holistic approach to standard setting in which panelists examine score profiles to determine cut points (see Perie & Thurlow, 2012; Zieky et al., 2008). The basis of the performance profile method is a collection of student score profiles, ordered on total score, showing how the student performed on each item. Rather than use a specific range finding and pinpointing process, panelists review the ordered profiles and identify the first profile that represents borderline performance between the two performance levels. All profiles with the same total score are considered in determining the cut point, and taken together, demonstrate multiple ways students may achieve the same total score value. Because profiles contain student scores for all items, the method is suggested for use with alternate assessments or other measures that contain a small number of performance tasks.

Contrasts with Other Methods

The Condensed Mastery Profile Method described in this paper draws from aspects of each of the above holistic methods. Like the Body of Work Method, the Condensed Mastery Profile Method makes use of range finding and pinpointing combined with logistic regression to identify cuts between performance levels. Consistent with both the Body of Work and Generalized Holistic methods, it implements two rounds of ratings in the process.

Despite drawing from other holistic methods, the Condensed Mastery Profile Method also differs in several key areas. Specifically, rather than booklets ordered on total score as used by the three aforementioned holistic methods, the Condensed Mastery Profile Method orders profiles of mastery using the total number of skills mastered. Furthermore, rather than items or performance tasks serving as the basis for setting standards, the Condensed Mastery Profile Method creates profiles of mastery based on probabilities of mastery obtained from a DCM for each skill measured by the assessment. Each profile is assigned a performance level, rather than the collection of item responses, or determining the borderline between two levels as is used in the Performance Profile Method. Furthermore, profiles for the Condensed Mastery Profile Method are selected from among the most frequently observed in the student data, rather than the highest and lowest papers within a range finding group, as in the Body of Work Method.

Condensed Mastery Profile Method

The Condensed Mastery Profile Method draws upon relevant holistic standard setting methods while leveraging the map structure underlying the diagnostic assessment. The process that follows provides a high-level description of the methods used for the standard setting event.

Each step in the Condensed Mastery Profile Method is described in the sections that follow. This includes steps prior to the standard setting meeting, such as specification of policy performance level descriptors, creation of mastery profiles, selection of panelists, and training. Following that, a discussion of the process for setting standards is provided, including rounds of range finding, pinpointing, statistical adjustment, review of impact data, and procedural evidence collection. Additional detail can also be found in the technical reports for each blueprint testing model (Karvonen, Clark, & Nash, 2015a; 2015b).

Performance Level Descriptors

The first step in the Condensed Mastery Profile Method involves defining performance categories. For the DLM assessment, policy performance level descriptors (PLDs) were developed by the assessment's stakeholders (consortium state education agency partners) to inform the interpretation of assessment results. The language of the PLDs was developed through a series of discussions spanning a six-month period using an iterative and consensus-based process. State partners began by reviewing language used in other state assessment systems and consortia. Partners then suggested and refined the language used to describe the levels, including soliciting local feedback within their states. All states participating in the consortium required four performance levels for score reporting and accountability purposes, thus requiring three cut points to be specified for each assessment. Upon reaching an agreement, the text of the PLDs was finalized, and is provided in Table 1. In contrast with many other standard setting approaches, grade and content-specific PLDs were not used during the standard setting process. Instead, grade and content PLDs emerged from the standard setting process using the mastery profiles and underlying map structure to describe the specific skills typically mastered by students in each performance level, rather than relying on stakeholder judgment. The final grade and content PLDs were included in score reports to support interpretation of results.

Insert Table 1 about here

Profiles of Student Mastery

Profiles of student mastery should be constructed consistent with the scoring method used for producing student score reports. Mastery profiles used in standard setting represent possible patterns of skill mastery as demonstrated in the population.

Mastery Thresholds. Mastery thresholds are applied to probability estimates obtained from the DCM to indicate whether the student is classified as a master or non-master of each skill measured by the grade and content area. Because DLM reporting is at the linkage level for each EE, a threshold for mastery was specified at the linkage level.

The selection of a mastery threshold (and the specific model used to obtain them) will likely be unique to each assessment program. For DLM assessments the threshold was based on data analysis and input from state partners and the consortium's Technical Advisory Committee (TAC). DCM posterior probabilities that are near 0.5 represent maximum uncertainty in whether the student is classified as a master or a non-master. Probabilities near 1.0 or 0.0 represent maximum certainty in mastery status. For DLM assessments, a value of 0.8 was selected as the mastery threshold to reduce the likelihood of measurement error impacting mastery classifications while accounting for the variability that might be expected in performance from students with significant cognitive disabilities.

Profile Selection. Student data from the assessment was used to determine profiles of mastery for all students who participated in the operational testing window prior to a cutoff date one month before the standard setting event. The file summarized the highest linkage level mastered for each EE for each student. The number of students who had profiles available for each grade, content area, and blueprint testing model combination ranged from 405 to 7,062.

Due to the number of EEs included on the test blueprints, and a student being able to master between zero and five linkage levels for each, there were a number of possible ways for students to demonstrate mastery for any given total linkage level value. A program was written using the *R* programming language (R Core Team, 2013) to select exemplar profiles to include in standard setting that also were substantially different enough to capture variations in performance. The program was written to select the three most common profiles of student mastery from the available profiles for each total linkage level value differing on at least three EEs. More detail on the selection of three common profiles is provided in the Discussion section.

The selection program was written to read in the student-level data file and select the three most commonly occurring mastery patterns for each total linkage level value. One caveat was introduced into the program to prevent the selected profiles from being overly similar, which could potentially negatively impact the standard setting process. For this reason, the program ensured that the three profiles selected were the most frequently occurring but also differed on the highest linkage level mastered for at least three EEs. As an example, the program selected the three most common ways to master 27 linkage levels over all EEs on the blueprint for a grade and content area, from among all observed patterns for mastering 27 linkage levels. However, if two of the returned profiles differed on only two EEs, the next most common profile was identified and retained. The resulting data set included three rows of mastery profiles for each possible total number of linkage levels, from one linkage level mastered up to the maximum number of total linkage levels mastered for each grade, content area, and testing model.

In some instances, three profiles were not available for every possible linkage level value due to patterns not being observed in the data. In these instances, the program returned the most common mastery patterns, up to the number of available profiles. Any remaining profiles, up to the three required, were created by test development teams. To create profiles, teams reviewed profiles at adjacent total linkage level values and created likely mastery profiles for any total linkage level values where additional profiles were needed. These rows were added to the original dataset output by the *R* program to produce a single file for each grade, content area, and testing model that contained the three profiles for each total linkage level value.

Profile PDF creation. A program was written using the *R* programming language (R Core Team, 2013) to produce profiles of student mastery in PDF format to be used at the standard setting event. These exemplar profiles of student mastery were created from the previously described dataset that contained the most common profiles at each total linkage level value. Figure 2 provides an example profile of student mastery. The profiles included a row for each EE on the blueprint, organized into larger conceptual areas. For each EE, the five linkage level descriptors were included on the profile, with shading in the cell to distinguish linkage levels that were mastered from those that were not. Each profile additionally included the total linkage levels mastered, an identification code, and the grade, content area, and model for which standards were being set.

[Insert Figure 2 about here]

Panel Creation

Consistent with other standard setting methods, the Condensed Mastery Profile Method relies on well-qualified panelists to set standards. State partners recruited individuals to volunteer as standard setting panelists. They sought individuals with content knowledge and those with expertise educating students with significant cognitive disabilities. Panelists were selected from the pool of volunteers, balancing breadth and type of experience with state representation across the panels. A total of fourteen panels were created, consisting of between four and eight members. In instances where representation isn't needed across multiple states, the use of four-six panelists may be desirable. Table 2 summarizes the standards each panel was responsible for setting. Because each assessment required three cut points, the panels were responsible for specifying 120 cut points for 40 assessments.

[Insert Table 2 about here]

Training

Panelist training was provided both in advance of and during the standard setting event. Advance online training consisted of familiarizing panelists with topics relevant to the assessment program, including information on students who take the assessment, the content and design of the system, a high-level overview of how student mastery of skills is determined and reported, and the process for setting standards using mastery profiles. Participants also completed an online quiz to help indicate areas of less comfort that could be covered and clarified during the on-site training.

The on-site training during the standard setting event consisted of a review of key topics as identified in the advance training quiz, as well as specific information about how to rate profiles. Facilitators also reviewed a notebook of available resources, which included hints for making ratings, diagrams of elements in the DLM system, and a glossary of terms. To familiarize panelists with the content of the grade and subject for which they were setting standards, they were also given node description booklets, a blank profile for annotating, and the test blueprint. A practice round of range finding was conducted at the conclusion of the training process to provide panelists the opportunity to practice rating mastery profiles, during which they first familiarized themselves with the grade-level content through review of the blank profiles and node descriptions.

Training was also provided for the panel facilitators and other support staff in advance of the standard setting event. An overview of the standard setting process was provided, along with a script detailing their role in the process. Facilitators, who were each assigned to a panel, were also given training on the Excel workbooks to be used at the standard setting event and given time to practice entering values during a mock range-finding event. This process also allowed for updates to be made to the script and agenda based on outcomes of the facilitator training.

Range Finding

The purpose of the range-finding process was to identify general divisions between performance-level categories. Two rounds of range finding were implemented to arrive at the general range where a cut between performance levels was likely to be. During both rounds, panelists referred to folders containing the exemplar student profiles along the full range of linkage levels mastered, in increments of five (e.g., five, ten, or fifteen levels mastered). For each of these total linkage level mastery values panelists were provided three exemplar profiles, showing the three most common patterns of mastery for obtaining that number of linkage levels, accounting for overly similar profiles.

During range-finding, panelists independently reviewed the contents of the profiles. When requested by the panelists, facilitators projected sample assessment items for each EE and linkage level. For each profile, panelists identified the performance level for each that best described the student's performance and recorded decisions on a rating sheet. When all panelists had completed their ratings, they shared the performance category assigned to each profile by show of hands. The facilitators recorded the ratings by entering these values in the Excel workbook, which was projected at the panel table. One panelist was assigned the task of verifying correct entry into Excel as values were added. When all values had been entered, the panelists discussed what information influenced their decision to categorize a profile to a certain performance level. Table facilitators encouraged conversation but did not otherwise contribute to the discussion or suggest panelists modify their ratings. Following discussion, panelists had the opportunity to revise their ratings during the second round of rating.

After all of the second round ratings were entered, logistic regression functions built into the Excel workbook identified the points of maximum uncertainty between performance levels. Specifically, logistic regression is used to find the value for which the probability of being classified into each of two contiguous categories is 0.5—which is the point of maximum disagreement (Kingston & Tiemann, 2012). Because the specification of cuts relies on the point of maximum disagreement between panelists, consensus on ratings for profiles was not needed. Results of the logistic regression were used to select pinpointing profiles. On-site psychometricians reviewed all workbooks prior to finalizing the range-finding cut points, and in instances where logistic regression did not provide a value (e.g., in instances where the panel had complete agreement), psychometricians visually inspected the results to identify the point of inflection between performance levels.

Pinpointing

Cut points identified during range finding were used to populate folders for the pinpointing process. Pinpointing folders for each cut (e.g., the cut between at Target/Advanced) included profiles with a range of seven total linkage levels mastered, plus three and minus three from the cut point identified during range finding. As an example, if the cut identified during range finding was 21, profiles for pinpointing were provided for the range of 18-24 linkage levels. Each linkage level mastery value had three available profiles, for a total of 21 profiles to be reviewed per cut. Profiles were ordered in each cut point folder from least linkage levels mastered to most linkage levels mastered.

Following the same procedure as range finding, panelists independently reviewed each profile and indicated the performance-level classification on a rating sheet. Ratings were shared with the group by show of hands and recorded in the projected Excel workbook, with one panelist confirming the accuracy of all values entered. Following discussion and the second round of ratings, the logistic regression function built into the workbook identified the most likely cut points based on panelist ratings. Psychometricians on-site for the standard setting event reviewed all final cut points, and in instances where the logistic regression function did not produce a value, (e.g., in instances where the panel had complete agreement), visually inspected the results to identify the point of inflection between performance levels.

Statistical Adjustment

Because the selected panelists represented a small sample of all possible experts, and some amount of variability in final cut points from the true value is to be expected should the process be repeated with a different panel, a statistical adjustment procedure was implemented. The adjustment borrows strength from data at other grade levels under the assumption that, barring information to the contrary, there is little or no reason to expect the percent of students in one grade to dramatically differ from the percent in a contiguous grade. In essence, cut points are smoothed over grades using a statistical rather than judgment-based procedure. The statistical adjustment was applied for each set of panel-recommended cut points (i.e., for each grade, content area, and testing model).

First, a frequency distribution of the number of students mastering each number of linkage levels was created with associated cumulative proportions. Next, a probit transformation was applied to identify the *z*-score associated with the cumulative proportion of students for each linkage level mastery value. *Z*-score values at the top of the distribution, where the proportion is equal to one, were defaulted to 3.5. Following this step, the *z*-score associated with each panel-recommended cut was identified. Weighted rolling averages were created for each cut, where the grade of interest was weighted 0.4, the contiguous grades weighted 0.2, and all other grades were weighted 0.1. Finally, using a table of probit-transformed cumulative proportions, the linkage level for the cut was identified, for which the *z*-score was closest to the weighted rolling average.

Impact Data

The role of impact data varies across standard setting processes. In the application of the Condensed Mastery Profile Method, the role of impact data was intentionally minimal, with content-based rationales guiding the recommendations provided by each panel. For evaluation purposes, the percent of students classified into each performance level was calculated for both the panel-recommended and statistically adjusted cut points. These values were shared with relevant stakeholders (the consortium's TAC and state partners) to aid in their decision-making process when determining the final cut points to be implemented for consortium-wide scoring and reporting purposes.

Grade and Content-Specific Performance Level Descriptors

Because the approach to standard setting described here relied on content-based judgments of student mastery profiles, grade- and content-specific PLDs were not developed or used for the standard setting event. Rather, the grade- and content-specific PLDs emerged as a result of the standard setting event, with student mastery profiles serving as the basis for their creation. Test development teams drafted the language for the grade- and content-specific PLDs using the test blueprint, the cut points from standard setting, sample mastery profiles, and other test development documentation used in the item writing process.

Following the drafting of grade- and content-specific PLDs, state partners reviewed and provided feedback for a subset of grades. Their feedback was incorporated into all documents, which then underwent a full editorial review prior to their release and incorporation into score reports to describe the knowledge, skills, and abilities typical of students classified into each performance level by grade, content area, and model.

Procedural Evidence

Evidence was collected from a variety of sources to support the final cut point determinations made using the Condensed Mastery Profile Method. Sources of evidence included convergence plots, standard errors of pinpointing ratings, and panelist feedback obtained from a survey at the close of the event.

Convergence

During range-finding and pinpointing, panelists gradually narrowed the range to identify the point where a cut between performance levels should be specified. Because of the use of logistic regression, consistency of ratings across panelists was not necessarily the desired outcome. Rather, the expectation was that panelist ratings would converge toward an increasingly narrow set of profiles to arrive at a final cut. To summarize the degree to which panelist ratings converged on a cut point value, box and whisker plots were created. Figure 3 is an example plot for ratings obtained from the ELA 9-10 grade band panel. The plots summarize the median, first and third quartiles, and the range of frequencies with which each total linkage level mastery value was classified into each performance level for each round of rating. These values can be compared to the final adjusted cut point values for each performance level.

[Insert Figure 3 about here]

The cut points represent the lowest value included in the higher performance level. For example, a cut point of 18 means that a linkage level mastery of 18 or greater is considered Approaching. Grade 9 and 10 are assessed as a single grade band for ELA in the integrated model. These convergence plots provide one source of evidence that the panel process worked as intended. The plots demonstrate that the ranges of profiles categorized into each performance level narrowed from round one to round two for the range-finding and pinpointing processes.

Standard Errors of Pinpointing Ratings

After the standard setting event, the standard error of the panelist pinpointing ratings was calculated using the frequency distributions from the panelists' final round of ratings. The values were computed by dividing the standard deviation of the frequencies of panelists' final pinpointing ratings by the square root of the number of total ratings. For all performance levels (*n*=160), the standard error values ranged from 0.08 to 1.25, with a median of 0.20. These findings indicate overall that the multiple rounds of both range finding and pinpointing resulted in a small amount of variability in the final pinpointing ratings.

Results from Panelist Evaluation

At the conclusion of each standard setting meeting, panelists were asked to evaluate the process via a survey. The survey asked for the panelists' feedback on the training provided, the process for setting standards, the professional benefits related to attending the standard setting meeting, and their overall feedback on the specific cut points. Items were presented on a Likert scale, ranging from strongly disagree (SD) to strongly agree (SA).

Overall, panelists provided strong support for the methods used to set standards. Panelists indicated that the training provided the information needed to complete tasks during the event. Panelists felt confident rating profiles and understood the knowledge, skills, and abilities each profile represented. In addition, panelists reported being satisfied with the cut points determined by their panel and were confident the meeting provided valid cut point recommendations. The complete set of survey results can be found in Karvonen et al., (2015a; 2015b).

Confidence in Panel-Recommended Cuts. Part of the survey data collection process included panelist feedback regarding their panel's final cut points. Panelists were asked to indicate by how much, if any, they would adjust their panel's final cut point value for each cut. Across all cut points specified by their panel, the vast majority of panelists (95%) indicated they would not adjust the cut point from the panel-recommended value. Within the 5% of cases where panelists recommended adjustments, in most instances the recommendation was for only one of the three cut points for the grade/subject/model, and the recommended change differed from the panel-recommended value by only one linkage level.

In addition to providing feedback regarding personal recommended changes to the panel-recommended cut, panelists were also specifically asked to indicate their comfort with their table's final panel-recommended cut points with a *yes* or *no* response on the survey. Across all panelists, panels, grades, and cut points (*N*=858), the vast majority of panelists reported comfort with the panel-recommended cut points (95.9%). Panelists indicated discomfort with the panel-recommended cut in only 4.1% of responses (*n* = 35). For 26 out of 40 panels (65%), panelists indicated complete comfort with all three panel-recommended cut points for the grade level.

Additionally, panelists were given the opportunity to indicate whether they would defend the panel-recommended cut points against the argument the cut points were set too high or too low. Table 3 presents the percent of panelists selecting each option, along with the number of panelists responding to the item. Taken together, the survey data pertaining to panelist comfort with the panel-recommended cut points indicates the profile-based approach leads to cut points that panelists support, would not modify, and would defend against criticism because of the content-based rationales for why profiles were categorized to their respective performance levels.

[Insert Table 3 about here]

Discussion

The Condensed Mastery Profile Method appears to be a useful and technically sound approach for setting performance standards for diagnostic assessment systems. DCM posterior probability estimates are subjected to a threshold to determine mastery or non-mastery of the attribute. The number of mastered attributes is summed to determine the total number of attributes mastered. The most common patterns of attribute mastery are identified and used to create profiles of student mastery that summarize attribute mastery patterns. Panelists then use the profiles of mastery to determine cut points between pre-determined PLDs.

The Condensed Mastery Profile Method draws from other holistic standard setting approaches (e.g., Cizek & Bunch, 2007; Kingston & Tiemann, 2012) to determine final cut points, including rounds of range finding and pinpointing to reduce the range until the cut for each performance level is identified. The procedures for selecting profiles to be included in the standard setting event address limitations common to holistic approaches to standard setting, including the presence of missing data or the inclusion of inconsistent patterns of performance, by selecting the three most common profiles for each linkage level mastery value and using content-based “simulated” profiles where necessary. Furthermore, the Condensed Mastery Profile Method is consistent with recommendations for best practice in the literature (e.g., AERA et al., 2014; Cizek, 2012).

The sources of evidence obtained from the standard setting event indicate that using the Condensed Mastery Profile Method to set performance standards for a diagnostic assessment resulted in cut points that panelists were confident about. Cut points represented panelists’ beliefs regarding fair delineations for categories of students that represent what they know and can do at each level.

Significance and Relevance to the Field

As DCM continues to grow in prevalence due to its ability to provide fine-grained reporting, standard setting methods must be developed to accommodate profiles of student mastery rather than typical methods based on item-level performance, scale score values, or student work products. The Condensed Mastery Profile Method described here is flexible enough to allow for differences in test administration based on the diagnostic assessment’s structure, as indicated by the implementation for two blueprint testing models and two content areas over multiple grade levels. While these differences resulted in varying numbers of total linkage levels mastered, the Condensed Mastery Profile Method successfully resulted in cut points for each grade, content area, and model that panelists felt comfortable with and were supported by content-based rationales for their values. In addition to the variations above, the method can also accommodate varying numbers of performance levels, can accommodate varying numbers of panelists per panel, and can be applied to both general and alternate assessment populations.

The method as described in this paper resulted from a series of decisions made throughout the process. Many of these decisions are flexible and can be adjusted based on the needs of the individual assessment program. For instance, the specific DCM used and threshold for specifying mastery should be determined based on the design of the assessment and student population. Additionally, researchers made decisions regarding the number of panelists to include and the number of profiles they reviewed per linkage level value. Three profiles per linkage level were used to demonstrate varying performance for the linkage level value while not providing too many profiles to review, as feedback from the TAC and a mock panel prior to the event indicated the process could become too complex if panelists were asked to do too much. However, this value might be adjusted in subsequent applications if fewer pinpointing profiles were provided (e.g., rather than seven points around the cut from range finding, five were used), fewer cuts were specified, or panels specified cuts for fewer grades during the meeting. Applications of the Condensed Mastery Profile Method should carefully consider whether similar decisions make sense given the purpose, design, and intended use of the assessment system.

Limitations and Future Research

As with any standard setting method, the process of specifying cut points can introduce error. Each panel of participants was selected from among the full body of volunteers. As such, the selection of different panelists may have resulted in the specification of a different set of cut points. Additional potential for error is further introduced in making dichotomous mastery status designations. Rupp, Templin, and Henson (2010) caution against instances where standard setting relies on a multi-stage approach to setting performance standards, whereby a total score on a latent trait is determined, and then that score is used to classify students to performance levels. They argue that reporting the EAP probabilities combat this issue, since they directly quantify the certainty in estimates. However, in operational applications of diagnostic measurement, state policies often dictate that performance levels must be specified in order to feed into state accountability metrics. The statistical adjustment procedure was implemented as part of the Condensed Mastery Profile Method to ameliorate the issue of measurement error, however, measurement error might still have an impact on student classification to performance levels.

Additionally, the purpose of using logistic regression is to determine the point at which there is maximum disagreement in the ratings among panelists to specify a cut point (i.e., the point at which the probability of being classified into either performance level is 0.5). Where panelists had complete agreement on all profile ratings, logistic regression failed to produce a cut point. In these instances, psychometricians on-site at the standard setting event visually inspected the results of the range-finding or pinpointing round and identified the point of inflection so the process could proceed. This approach is similar to the method of identifying the median of the panelists' ratings in instances where logistic regression cannot be employed as identified by Morgan and Michaelides (2005). Future research should investigate alternative approaches to identifying cut points that do not rely on logistic regression to arrive at the cut point. This issue may also be combatted by creating larger panels so there is less likely to be complete agreement.

Due to these limitations of using logistic regression to identify cut points, the researchers borrowed from other methods (e.g., Kingston & Tiemann, 2012) by calculating the standard error of pinpointing values. However, these values were also limited based on the range of profiles available to panelists in each performance level. The final standard error values were highly contingent on the range of linkage levels evaluated for each performance level, rather than purely representing variation in panelist ratings. Future studies should consider alternate ways to report panelist variability in ratings.

An additional area for future research to expand upon the Condensed Mastery Profile Method would be to identify an alternative to condensing the profiles prior to obtaining judgments from panelists. Instead, a wide variety of profiles could be presented to panelists for classification and a method derived to create a decision rule for specifying the cut points from among all the ratings.

As with any standard setting method, the Condensed Mastery Profile Method benefits from instances where a wide range of total linkage levels are available. The process was impacted in instances where the total number of linkage levels available was low due to a narrow blueprint. Pinpointing ranges can overlap when the range is too narrow resulting in a restricted range upon which to identify multiple cut points. While the procedures and evidence described here were applied to one testing program, the method may be generalized to other diagnostic assessment programs that require performance categories for accountability or other purposes.

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.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understanding of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*, 2-14.

Broaddus, A. (2012, April). Modeling understanding of foundational concepts related to slope: An application of the attribute hierarchy method. In J. Leighton (Chair), *Cognitive diagnostic assessment: Lessons from practice.* Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Cizek, G. J. (2012). *Setting performance standards: Foundations, methods, and innovations.* New York, NY: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: SAGE Publications.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement, 6,* 263-268.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan's *Educational measurement* (4th Ed., pp. 433-470). Washington DC: American Council on Education.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY: Cambridge University Press.

Karvonen, M., Clark, A. K., & Nash, B. (2015a). *2015 integrated model standard setting: English language arts and mathematics* (Technical Report No. 15-02). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from http://dynamiclearningmaps.org/about/research/publications

Karvonen, M., Clark, A. K., & Nash, B. (2015b). *2015 year-end model standard setting: English language arts and mathematics* (Technical Report No. 15-03). Lawrence, KS: University of Kansas, Center for Educational Testing and Evaluation. Retrieved from http://dynamiclearningmaps.org/about/research/publications

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 218-248). Mahwah, NJ: Erlbaum.

Kingston, N. M., Karvonen, M., Bechard, S., & Erickson, K. (2016). *The philosophical underpinnings and key features of the Dynamic Learning Maps Alternate Assessment*. Teachers College Record (Yearbook), 118(14). Retrieved from http://www.tcrecord.org

Kingston, N. M., & Tiemann, G. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 201-224). New York, NY: Routledge.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Morgan, D. L., & Michaelides, M. (2005). *Setting cut scores for college placement* (Research Report No. 2005-9). New York, NY: The College Board. Retrieved from https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2005-9-setting-cut-scores-college-placement.pdf

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Perie, M., & Thurlow, M. (2012). Setting achievement standards on assessments for students with disabilities. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 347-377). New York, NY: Routledge.

R Core Team. (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-262.

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

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239-257.

Skaggs, G., Hein, S. F., & Wilkins, J. L. (2016). Diagnostic profiles: A standard setting method for use with a cognitive diagnostic model. *Journal of Educational Measurement, 53*, 448-458. doi: 10.1111/jedm.12125

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing, 11*, 1-23. doi: 10.1080/15305058.2010.518261

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317-339.

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48,* 165-187. doi: 10.1111/j.17453984.2011.00142.x

Zieky, M., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Table 1

Text of the Performance Level Descriptors

Performance Level Descriptors
The student demonstrates *emerging* understanding of and ability to apply content knowledge and skills represented by the Essential Elements.
The student's understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements is *approaching the target*.
The student's understanding of and ability to apply content knowledge and skills represented by the Essential Elements is *at target*.
The student demonstrates *advanced* understanding of and ability to apply targeted content knowledge and skills represented by the Essential Elements.

Table 2

Standards Set by Each Panel

Panel	Assessment Grades/Courses
ELA IM	3, 4, 5
ELA IM	6, 7, 8
ELA IM	9-10 grade band, 11-12 grade band
ELA YE	3, 4, 5
ELA YE	6, 7, 8
ELA YE	9, 10, 11
ELA YE	English 2, English 3
Math IM	3, 4, 5
Math IM	6, 7, 8
Math IM	9, 10, 11
Math YE	3, 4, 5
Math YE	6, 7, 8
Math YE	9, 10, 11
Math YE	Algebra 1, Algebra 2, Geometry

IM = integrated blueprint model assessments; YE = year-end model assessments

Table 3

Percent of Panelists Selecting Each Response Option During Standard Setting Evaluation

Question	**SD**	**D**	**A**	**SA**	***n***
1. I would defend the group's At Target decisions against criticism that they are too high.	0	1	44	55	99
2. I would defend the group's At Target decisions against criticism that they are too low.	0	2	39	59	99
3. I would defend the group's Advanced decisions against criticism that they are too high.	0	1	41	58	98
4. I would defend the group's Advanced decisions against criticism that they are too low.	0	1	40	59	99
5. I would defend the group's Approaching Target decisions against criticism that they are too high.	0	2	43	55	98
6. I would defend the group's Approaching Target decisions against criticism that they are too low.	0	2	40	58	99