

AN EXPERIMENTAL STUDY OF STANDARD SETTING METHODS FOR DIAGNOSTIC PROFILES

by

ZACHARY R. FELDBERG

(Under the Direction of Laine Bradshaw)

ABSTRACT

Cognitive diagnostic models (CDMs) provide pedagogically relevant information in the form of a student profile of multiple binary categorizations of students into mastery or nonmastery statuses on latent traits called attributes. Federal educational accountability requires state accountability measures to designate students into one of at least three ordinal levels of proficiency. To bridge the gap between educational accountability testing and pedagogical relevance, CDMs have been proposed for use as state accountability measures. This study examines variables related to the methods used to map the multiple binary categorizations of the students profiles onto the federally required ordinal levels of proficiency, a process called standard setting. We conducted an experimental study of standard settings for DCMs by asking five panels of experts to categorize student profiles into levels of proficiency. We manipulated the number of profiles panelists viewed, the variability of the set of profiles panelists viewed, and the range of attributes mastered across the sets of profiles panelists viewed to determine if these factors impact the resulting cut point. Results indicate that the methods of standard setting are resilient to changes in the number of profiles that panelists view or the variability of profiles panelists view but are effected by the range of profiles panelists view, indicating that while

standard settings produce consistent results they lack objectivity. As arbitrary distinctions in student performance, we point out that it is difficult to justify their continued use in educational measurement.

INDEX WORDS: Cognitive Diagnostic Models, Standard Setting, Educational Policy,
Educational Measurement, Educational Improvement

AN EXPERIMENTAL STUDY OF STANDARD SETTING METHODS FOR DIAGNOSTIC
PROFILES

by

ZACHARY R. FELDBERG

B.S., The University of Georgia, 2011

M.A.T., The University of Georgia, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2023

© 2023

Zachary R. Feldberg

All Rights Reserved

AN EXPERIMENTAL STUDY OF STANDARD SETTING METHODS FOR DIAGNOSTIC
PROFILES

by

ZACHARY R. FELDBERG

Major Professor:	Laine Bradshaw
Committee:	George Engelhard
	Laura Lu
	Matthew Madison

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2023

DEDICATION

To all the students that have been told that did not meet expectations, you too can get a PhD. For all the teachers that work hard with the hardest to teach students but are told their work is not proficient. You are not defined by arbitrary, socially constructed labels of proficiency created by disconnected academics and capitol hill pencil pushers.

ACKNOWLEDGEMENTS

I would first like to thank my advisor and committee chair, Dr. Laine Bradshaw, whose support was instrumental in this research. Her passion for improving teaching and learning knows no bounds. I will forever respect and try to emulate her focus, professionalism, and work ethic. I am grateful for the time and opportunities she has provided me over my academic career and hope to make her proud.

I would also like to thank the numerous other individuals who have provided guidance over the years. Thank you to my committee members for their assistance and direction in this research. I would particularly like to thank Dr. Matt Madison for his direction and support when times were difficult.

Finally, and most importantly I would like to thank my family and friends for their love, support, and encouragement over the years. I would like to particularly thank my parents for seeing the potential in me and telling fourth grade Zack that just because the test said I was not proficient did not mean I could not go on to teach graduate level statistics and apply statistics in real world applications. Not sure those were your exact words but doubtlessly it was your support that got me to where I am today.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xii
Chapter 1. Statement of the Problem	1
Federal Policy Requirements: Single Summative Determination	1
The Problems with Current Accountability Testing	4
CDMs: A Potential Solution	4
Obtaining a Single Determination of Student Performance with CDMs.....	5
Chapter 2. Literature Review	7
IRT-based Standard Setting Approaches	7
DCM-Based Standard Setting Approaches.....	12
Standard Setting Approach Proposed in Georgia’s IADA Application.....	22
Implications of Previous Methods for a DCM Standard Setting Approach	24
Research Questions	25
Chapter 3. Study Design	29
Establishing Conditioning.....	29
Panelists Selection Criteria	34

Panelists Recruitment.....	34
Creating the Profile Cards and Panelists Packets	35
Training Panelists: In-Advance Training.....	36
Training Panelists: Introduction and Before Standard Setting Training.....	38
Standard Setting: Range Finding	38
Standard Setting: Pinpointing	41
Independent Evaluation and Exit Survey.....	42
Post Standard Setting Facilitator Notes	42
Pilot Study.....	43
Data	44
Instruments.....	45
Analyses.....	45
Chapter 4. Results	53
Panelists Recruitment.....	53
Panelists Selection	53
Post In-advance Training Survey.....	56
Data Collection	58
Standard Setting Recommended Cuts Points.....	62
Change and Convergence of Rating	64
Standard Deviations and Standard Errors	69

Nonparametric Longitudinal Factorial Analysis.....	71
Panelist Exit Survey Ratings.....	78
Chapter 6. Discussion	79
Discussion by Question.....	79
Conclusion	86
Study Limitations.....	90
Future Research	91
References.....	93
Appendix A.....	99
Appendix B.....	101
Appendix C	103
Appendix D.....	105
Appendix E	109
Appendix F.....	110
Appendix G.....	113
Appendix H.....	114
Appendix I	133
Appendix J	134
Appendix K.....	136
Appendix L	137

Appendix M	138
Appendix N	139
Appendix O	142
Appendix P	144

LIST OF TABLES

	Page
Table 3.1: Prevalence of Each Profile Viewed by Panelists in Condition 2	33
Table 3.2: Georgia Performance Level Descriptors	37
Table 3.3: Example Coding of Panelists Ratings.....	40
Table 3.4: Table of Data of this Study	44
Table 3.5: Table of Instruments of this Study.....	45
Table 4.1: Panelists' Experience and Familiarity with Standards by Condition Group	55
Table 4.2: Results of Post In-Advanced Training Survey	56
Table 4.4: Set of Profiles Presented to Panelists During Range Finding Rounds	58
Table 4.5: Results of the Logistic Regression Functions Conducted After the Second Round of Range Finding.....	60
Table 4.6: Set of Profiles Presented to Panelists During Pinpointing Rounds	61
Table 4.7: Final Cut Points Produced by Logistic Regression Functions by Condition Group and Round	63
Table 4.8: Interrater Agreement of Profile Categorizations for Each Condition and Round	65
Table 4.9: Condition and Panelists Rating Changes by Round	67
Table 4.10: Cut Points and Standard Deviations of Cut Point for Final Round Ratings.....	69
Table 4.11: Standard Deviations and Standard Errors of Total Mastery Level for Final Round Ratings by Proficiency Level and Condition	70
Table 4.12: Results of Nonparametric Factorial ANOVA for Conditions 1, 2, and 3.....	73

Table 4.13: Results of Nonparametric Factorial ANOVA for the Middle Cut Point	74
---	----

LIST OF FIGURES

	Page
Figure 2.1: Weighted Distribution and Cut Score for a Single Panelists.....	16
Figure 2.2: Example Profile Used in the Condensed Mastery Profile Standard Setting	19
Figure 3.1: Example Convergence Plots.....	47
Figure 4.1: Final Achievement Level Ranges by Round and Condition Group.....	64
Figure 4.2: Relative Marginal Effects of Treatment Conditions for the Middle Cut Point Ratings.....	#

Chapter 1. Statement of the Problem

Current federal education accountability policy allows states to create new, innovative assessment systems under the Innovative Assessment Demonstration Authority (IADA) but requires those systems to produce a single summative designation of student proficiency. This presents a challenge when the new assessment systems, such as those based on cognitive diagnostic models (CDMs; Leighton & Gierl, 2007), measure students on multiple dimensions instead of one. Most legacy accountability systems are based in item response theory (IRT; Hambleton et al., 1991) frameworks, which produce a single score on a continuous latent scale. CDMs instead create profiles of student ability on multiple, dichotomous latent variables called attributes. Whereas an IRT test might describe a student's ability on a continuous scale from 500 to 600, a CDM would describe a student as master or nonmaster of a set of dichotomous attributes. CDM based assessment results may be preferable to IRT test results because they can provide pedagogically relevant information by indicating when students need more instruction on a specific attribute or are ready to move on. For CDM-based innovative assessments systems to meet current federal accountability mandates, however, a method is needed for making a single proficiency determination from the multidimensional student attribute profiles.

Federal Policy Requirements: Single Summative Determination

The bulk of current federal education accountability policy is encapsulated in the 2015 Every Student Succeeds Act (ESSA; P.L. 114-95). This legislation is a reauthorization of the Elementary and Secondary Education Act (ESEA; P.L. 89-10), a half-century-old act aimed at

providing equal education opportunities to all students by providing funds to low-income schools. Since its inception, ESEA has required the regular evaluation of programs receiving funds (Resnick, 1980). In the 1994 reauthorization of ESEA as the Improving American Education Act (IASA; P.L. 103-382), Congress shifted from requiring program evaluation to establishing a system of accountability by mandating that states develop academic content standards and tests aligned to those standards (Shepard, 2008). These requirements were expanded in the 2001 reauthorization of ESEA as No Child Left Behind (NCLB; P.L. 107-110). In order to receive Title I funds, NCLB required states to test students yearly in grades 3 through 8 and once more in high school in both mathematics and reading or language arts and once in science in each grade band between grades 3 through 5, 6 through 9, and 10 through 12 (NCLB § 1111 (b)(3)(C)(v)(I-II)). In 2015, NCLB was reauthorized as the Every Student Succeeds Act (ESSA), which reformed NCLB by providing states with greater autonomy and flexibility in creating their own accountability plans, but largely kept the testing requirements intact with slight alterations.

The state tests mandated under ESSA must make a single, summative determination of student achievement among at least three levels of achievement for each student. The law states that

Each State, in the plan it files ... shall provide an assurance that the State has adopted challenging academic content standards and aligned academic achievement standards (referred to in this Act as ‘Challenging State academic standards’), which achievement standards shall include not less than 3 levels of achievement, that will be used by the State, its local educational agencies, and its schools to carry out this part. (ESSA § 1111 (b)(1)(A))

In practice, the “3 levels of achievement” require students to be designated as students that do not meet, meet, or exceed proficiency expectations, though some states have more than three levels, the labels differ from state to state, and there is no established definition or meaning of

proficiency. For example, Florida has five levels, labeled “Level 1” through “Level 5” whereas Georgia uses four levels called “Beginning Learners,” “Developing Learners,” “Proficient Learners,” and “Distinguished Learners.” What it means to be proficient is not clarified at the federal level and varies among states (Linn, 2005).

The use of proficiency designations are tied to ubiquitous IRT based testing frameworks and has significant effects on the development of accountability tests. When developing such systems under such frameworks, policymakers must first choose the number and name of the “not less than 3 levels of achievement,” which are commonly called performance levels or achievement levels. Once the name and number of performance levels is decided, policy makers then define general performance level descriptors (PLDs) for each (Perie, 2008). The number of levels and the PLDs themselves are inherently arbitrary policy decisions (Linn, 2005).

The federal government requires the development of policy PLDs, which define what students at each proficiency level are expected to know or can do. Once the policy PLDs are developed, policymakers, content specialists, and or learning experts develop the “challenging academic content standards” required by ESSA for each subject area that is to be tested (Perie, 2008). Then, using the policy PLDs and standards, test items have traditionally been developed to measure a unidimensional, continuous latent construct of student ability or achievement under an IRT based framework. These tests usually produce a raw score that is transformed into a scaled score for interpretability and equating purposes. Test developers then establish cut scores along the continuum of the scaled scores using standard setting methods, such as the popular Angoff Method (Angoff, 1984), Modified Angoff Method (Hambleton & Plake, 1995), or Bookmark Method (Mitzel et al., 2001). Despite the term “standard setting,” they are really cut

score setting methods that define the boundaries between different levels of achievement (Ricker, 2006).

The Problems with Current Accountability Testing

While standard setting methods to define achievement levels have long been used for accountability purposes within summative, high-stakes testing situations, the achievement levels provide little instructionally relevant information. Telling a teacher that a student achieved “proficient” on a test does little to explain what specifically that child can do or needs to learn, even with detailed PLDs. Further, the tests are generally given as summative, end of grade or end of course tests, so there is little time for teachers to receive and act on the results of the tests. This is partially due to the inherent limitations of IRT based tests which are designed to be summative measures that rank order students on general knowledge, skill, or ability constructs, but are less well suited for providing detailed feedback. Educational experts have long known these limitations and called for the use of new methods of assessment that could provide detailed, pedagogically relevant, and timely information (Shepard et al., 2018; Wilson, 2018; Koretz, 2017).

CDMs: A Potential Solution

Under section 1204 of ESSA, called the Innovative Assessment and Accountability Demonstration Authority (IADA; ESSA § 1204), Congress gave the Federal Department of Education the authority to choose up to seven states to pilot innovative assessment models that could remedy some of the above-mentioned shortcomings. These assessments may include:

- (1) Competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year end assessments, or performance-based assessments that

combine into an annual summative determination for a student, which may be administered through computer adaptive assessments; and

(2) Assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs. (ESSA § 1204 (a))

Several states have been granted authority to test new assessment systems under IADA, including one testing program in Georgia called Navvy that utilizes CDM-based assessments (Bradshaw, 2017) . The Navvy approach exemplifies many of the innovations noted by the IADA. Navvy assessments are competency-based and are more easily embedded into instruction because they can be delivered in shorter, standard-level chunks throughout the course. Further, the CDM-based Navvy assessments allow students to demonstrate mastery on specific academic content standards and allow for differentiated student support based on students' individual learning needs.

For Navvy, or new systems that may be developed to use CDMs in the future, to meet the accountability requirements, a method for creating a single determination of student performance will have to be utilized.

Obtaining a Single Determination of Student Performance with CDMs

The main problem this study attempts to solve is how to obtain a single determination of student performance from a CMD based assessment system. This problem can be thought of as a challenge of dimensionality reduction. CDMs produce multidimensional results, often with several dozen attributes. How should these dichotomous, multidimensional results be mapped onto the one-dimensional, ordered scale of the performance levels?

To date, two studies have attempted to set cut points on a single dimension based on multidimensional diagnostic profiles (Clark et al., 2017; Skaggs et al., 2016). These studies have

both utilized similar approaches to standard setting methods utilized in an IRT framework. In both cases, they present a panel of experts with samples of student's multidimensional profiles of attribute mastery and ask the panelists to categorize each profile into a performance level. Their approaches raise important questions that will be the focus of this study and are discussed in the following chapter.

In this study, we systematically alter the process of profile selection, the variability of student profiles within a panel, and the range of profiles presented to panelists during a CDM standard setting. The variation of these factors is used to explore their impact on resulting cut points, level of agreement among panelists, and panelists opinions of the fairness and appropriateness of the standard setting process. This will add to the CDM and standard setting literature by examining the influence of profile selection, variability, and range on the standard setting process. Chapter 2 presents a review of methods used for standard setting with IRT and DCM-based assessments and then presents the research questions for this study. Chapter 3 provides methods for addressing this study's research questions.

Chapter 2. Literature Review

Standard setting is utilized to provide classifications of examinees based on expert judgments and assessment results, but not provided directly by the assessment result. We review methods for standard setting for both IRT- and DCM-based assessments in this chapter. For IRT-based methods, we focus on the most widely used approaches including the Angoff Method, the modified Angoff Method, and the Bookmark Method. We then detail the two approaches that have been developed for DCM-based assessments, namely the Diagnostic Profiles Method (Skaggs et al., 2016) and the Condensed Mastery Profile Method (Clark et al., 2017). Next, it discusses the CDM-based approach approved in Georgia under the IADA. This system is the Navvy assessment system created by Navvy Education, LLC. It then discusses aspects of the reviewed IRT and CDM methods for diagnostic profiles standard settings. It concludes with the research questions of this study that stem from the problem statement of the previous chapter and literature review.

IRT-based Standard Setting Approaches

The most widely used standard setting approaches in IRT frameworks are the Angoff Method with various modifications (Angoff, 1984; Ricker, 2006) and the Bookmark Method (Buckendahl et al., 2002; Lewis et al., 2012; Karantonis & Sireci, 2006).

The Angoff Method and Modified Angoff Methods

In the Angoff Method (Angoff, 1984), panelists are first asked to conceptualize a ‘minimally proficient student,’ that is just barely within the performance level (i.e., for ‘proficient,’ a student that is minimally proficient, for ‘exceeds proficiency,’ a student that minimally exceeds proficiency, etc.). The panelists are then asked to examine each item on a test and estimate the probability of correct response, from 0 to 1, for a randomly selected minimally proficient student. The estimates are summed to form the cut score for each judge, and then the averages of these summed cut scores define the final proposed cut scores. The proposed cut scores end the standard setting process, but do not finalize the cut score locations: Proposed cut scores are reviewed for policy purposes, and adjusted as needed, by policy makers.

When Angoff proposed this method in 1971, he included a footnote that provided a variation to his original method (Ricker, 2006). This variation, along with numerous other variations that aim to increase the agreement among the judges, have come to be known as the modified Angoff Methods. Ricker (2006) describes four types of modified Angoff Methods. First is the modification described in Angoff’s original paper, in which the panel is asked how many out of one hundred minimally proficient students would correctly answer each item, and then the percent is converted to a p -value. This approach is used with the thought that percentages of a group of students are easier to conceptualize than probabilities for a single student. The second method is to use an iterative process in which the panelist conduct two (Chang, 1999) or three (Busch & Jaeger, 1990) rounds of making judgements, discussing areas of disagreement, and then making judgements again (Hambleton & Plake, 1995). The successive rounds are meant to increase panelist agreement. A third modification is to provide the judges with normative data (Busch & Jaeger, 1990). Under this procedure performance data from a sample of students,

usually in the form of item difficulties, are presented to the panel of judges after an initial round of judgement. The judges modify their original judgments based on the normative data and then results are averaged. The fourth type of modification is for panelists to provide probabilities of correct response as well as apply relative weights to scores of subdimensions of the test (Hambleton & Plake, 1995). In this case, the cut scores are for multidimensional, polytomously scored test exercises and raters are asked to both set cut scores between achievement levels and to provide weights of importance for each dimension on each exercise. Hambleton and Plake (1995) used a complex performance assessment of professional teaching standards, where the performance tasks were polytomously graded on five dimensions. Raters were asked to set cut scores for each polytomously scored task and to weight the five dimensions.

Several authors have been critical of the Angoff Methods. They have found that the requirement for panelist to make judgements about each item on the test is labor intensive (Shepard et al., 1993) and the cognitive complexity of the tasks (imagining a minimally competent student and deriving the probability of their correct response) is too great (Pellegrino et al., 1999). Further, studies have found that the expert judges cannot accurately and consistently make the probability judgement for minimally proficient student success (Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009; Goodwin, 1999).

The Bookmark Method

The Bookmark Method (Lewis et al., 1996, 2012) was developed in large part to decrease the labor intensity and cognitive complexity of the standard setting process of the Angoff Methods. It begins with test developers creating items and providing them to a sample of examinees. Then the standard setting facilitators provide a set of expert judges with an Ordered

Item Booklet (OIB), a booklet containing each item of the test on a single page, ordered by difficulty from easiest to hardest according to the IRT difficulty parameter estimated from the sample of examinees. The judges are asked to independently start at the beginning of the booklet and place bookmarks between the pages of the items for which they believe a minimally proficient student's probability of success drops below a specified value, called the response probability (RP). Judges are sometimes given the opportunity to discuss disagreement of placement with other judges and move their bookmarks before final bookmark locations are collected. Lastly, cut scores for each performance level are set using the parameters of the items at or nearest the average cut score location of the judges. Then, using linking and equating approaches cut scores can be expanded to items or test forms not seen by the judges (Karantonis & Sireci, 2006).

As with the Angoff Method and Modified Angoff Methods, the Bookmark Method suffers from several limitations and challenges, including a lack of internal consistency, the problem of dealing with guessing, the challenge of item disordinality, and difficulties in justifying the RP. Each of these problems is discussed below.

Clauser and colleagues (2017) examined the internal consistency of judgements made in bookmark standard setting by providing experts with OIBs with varying difficulty. They found that the resultant cut scores were influenced by the difficulty of the respective booklets, indicating a lack of internal consistency and that the standard setting in essence takes place during item writing. This issue is explored in a DCM setting through the third research question (discussed at the end of this chapter) of this study.

The probability of guessing can also impact the results of standard setting using the Bookmark Method (Baldwin, 2021). The method explicitly calls for instructing judges to place

the bookmarks at locations where a minimally proficient student could correctly answer without guessing. Since some examinees will guess the correct answer, the method proposed by Lewis and colleagues includes a procedure for correcting for guessing by factoring it out of the model used to create the OIB. Baldwin (2021) found that this approach overcorrects, resulting in cut scores that will be too high. This issue is not applicable to a CDM standard setting because the panelists are viewing the student attribute profiles rather than specific items of the test.

Another issue that impacts the Bookmark Method is item disordinality, which occurs when the panelists disagree with the ordering of the OIB. Skaggs and Tessema (2001), found that this may be due to variations in the samples used to develop the OIB, differences in local curriculum among the judges, and the difficulty judges have with accurately estimating item difficulty. As with the problem of guessing, this issue is less applicable in a CDM context because panelists are not viewing specific items. On the other hand, panelists may see some attributes as more important than others. If this is the case, they may judge two profiles that have the same number of attributes mastered to represent different levels of proficiency. This issue is explored in the second question of this study discussed below.

A final issue related to the Bookmark Method is the selection and justification of an RP (Karantonis & Sireci, 2006). The RP is the probability that a student performing at the borderline of a performance band will correctly answer an item. It has generally been set to .67 and denoted by RP67, indicating a two-thirds probability of correct response for a minimally proficient individual. The RP value is important because it determines the scale location for the items within the IRT model and therefore the ordering of the items within the OIB. The arguments for using the RP67 is that that value is consistent with establishing mastery as compared to an RP50 because a higher RP value reduces the judges' perceived probability that a student that has not

mastered the material can correctly answer the question (Mitzel et al., 2001). Furthermore, when guessing is removed, RP67 provides an optimum decision rule by establishing the point where the item information function is at a maximum (Huynh, 2000). That said, the selection of the RP can change the ordering of the OIB and research has also found that an RP of .50 is defensible (Karantonis & Sireci, 2006).

Because the student profiles are presented instead of items, the RP is not as significant in a CDM context. Instead, the mastery threshold takes on a similar role. CDMs produce probabilities of mastery of each attribute where values from 0 or 1 represent greater confidence in nonmastery or mastery status respectfully. Values closer to .5 represent the greatest level of uncertainty of a student's mastery status. While .5 is often used as the mastery threshold, assessment developers or policymakers wishing to decrease the probability a student is deemed to have mastery when they in fact do not, may increase the probability. This issue is not specifically addressed in this study.

DCM-Based Standard Setting Approaches

To date, there have been two standard setting methods developed for use with DCMs. The first is the Diagnostic Profiles (DP) standard setting method developed by Skaggs and colleagues (2016) and the second is the Condensed Mastery Profile Method by Clark and Colleagues (2017).

Diagnostic Profile Standard Setting Method

The DP method was first implemented by Skaggs, Hein, and Wilkin (2016) to set the standard for a test of student readiness for 9th grade Algebra I. The authors simulated the results

of a DCM-based assessment using the responses from items from six forms of the *Trends in International Mathematics and Science Study (TIMSS) Assessment 2003 for Grade 8 Mathematics*. They used the six test forms to create a test bank of approximately 127 items with 1500-2000 responses each. Then, they defined four main content domains and fifteen finer grained attributes measured by 120 items. After defining the Q-matrix with these items and attributes, they used the compensatory reparametrized unified model (C-RUM; Hartz, 2002) and an additive approach to estimate item parameters. They first estimated the model with the four main content domains as the attributes in the DCM. They then iteratively divided the attributes into more fine-grained attributes, adding one attribute at a time and estimating the model again after each additional attribute. They found that the model with ten attributes maintained a stable solution without substantially increasing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) or decreasing the item root mean square error of approximation (RMSEA) below 95%. This approach resulted in a ten-attribute test with 115 items. From this, they drew 40 items to define an algebra readiness test, with each attribute measured by three to ten items. They reduced the number of items so that they could compare the DP Method to an Angoff standard setting.

Fifteen standard setting panelists were given an orientation and an opportunity to take the test and evaluate the items. They then discussed the following policy performance level descriptor, “A student is ready to take Algebra I if he or she has acquired sufficient mathematical knowledge and skill to be able to perform well enough in Algebra I to pass the course with a grade of at least a C” (p.452). It is worth pointing out that this PLD has a specific performance outcome (success in an Algebra I) which may not always be present in PLDs. During the orientation, panelists discussed which of the ten attributes combinations would likely indicate

readiness for high school Algebra I. They also participated in a practice standard setting by viewing five attribute profiles, making judgements, and then discussing how they made those judgements.

For the standard setting, facilitators selected 80 attribute profiles to present to judges based on their prevalence among the sample data, while ensuring to also include an all-mastery and all-nonmastery profile. These 80 attribute profiles accounted for approximately 75% of all profiles that were observed in the student sample. When considering this approach to profile selection, it is likely that the methods used to select the DCM specifications (number and nature of attributes) increased the data fit and explains why 80 profiles represent so much of the data. As described above, the attributes were selected in a way to maintain a stable outcome with regards to AIC and BIC fit indices. The test was not developed for a DCM a priori, but instead a DCM assessment model was retrofitted to data from a test designed for IRT, and the selected items and Q-matrix design were developed to fit the data. DCMs have 2^α possible profiles, where α is the number of attributes. This means Skaggs and colleagues' (2016) 10 attribute assessment would have 1024 possible profiles. 80 profiles represent 7.8% of all possible profiles. It is possible that results of a DCM based assessment that has more attributes or that was developed as a DCM a priori could have a greater variety of attributes profiles across examinees, in which case 80 profiles would not represent 75% of all profiles and potentially substantially less. For example, preliminary analysis of results of the Navvy assessment system, a functioning CDM-based assessment system designed to be a CDM a priori, suggest that the high representation of the 80 profiles is likely due to the number of attributes used and the process Skaggs et al. (2016) used to emulate a DCM based assessment. The preliminary analysis of a Navvy assessment that included sixteen attributes and 1098 examinees found that the 80 most

prevalent attribute profiles represented only 359 examinees or 32% of all profiles (compared to 75% for Skaggs et al.) and only .12% of all possible profiles (80 of 65,526 potential profiles where $2^{\alpha} = 2^{16} = 65526$). These aspects of profiles selection become important if standard settings facilitators want to provide panelists with a representative sample of student attribute profiles. The relationship between the representativeness of attribute profiles is in part what motivated the first and second questions discussed at the end of this chapter.

According to the authors, the attribute profiles were delivered to panelists with the all-nonmastery profile first, the all-mastery profile last, and the remaining profiles presented in random order. The authors noted that after judging approximately 20 attribute profile, panelists reported having developed a personal strategy for identifying which attributes were essential for students and which functioned in a compensatory way with other attributes.

The standard setting was carried out in two rounds. In the first round, panelist independently judged whether each profile met the PLD. This was a yes/no judgment. They also wrote descriptions of the strategies they used to make their ratings. Then on a second day, panelists were shown the results from the first round with each profile ranked by the number of proficiency designations. The panelists were asked to discuss the ratings from the previous day starting with the profiles that showed the greatest variability between proficiency and nonproficiency ratings. After discussion, panelist conducted a second round of judgements. Agreement on the ratings was found to be high for the majority of the profiles (above 80%) with only seven profiles in the first round and five profiles in the second round having agreement of less than 60%. The authors describe the personal strategies described by panelists in writing after the first round and during discussion before the second round as complex. These strategies highlighted the attributes panelists deemed essential (for examples, integers, fractions, decimals,

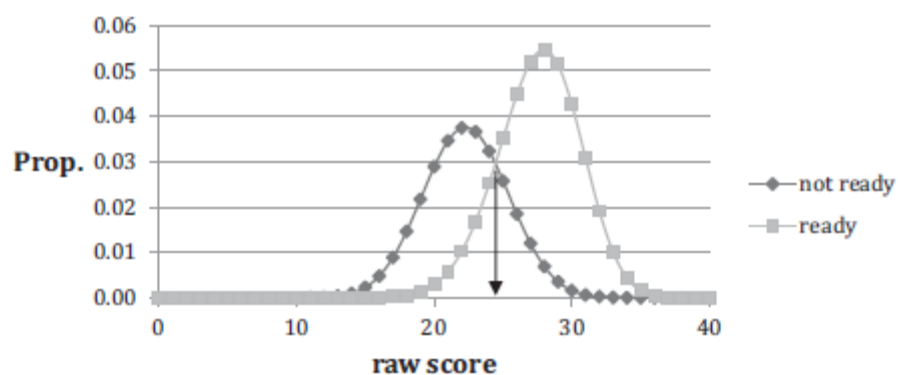
precents, and operations with fractions) to meet the PLD criteria and which were less essential (for examples, shapes and figures). The importance of some attributes over others likely reflected the specific nature of the PLD for algebra readiness.

Cut scores for the DP method were calculated by estimating the raw score distributions of each profile based on item parameter estimates and adapting a recursive formula for a compound binomial which predicted if a student would get an item correct based on their attribute mastery status. The raw score estimates were weighted by the proportion of examinees in each attributes profile. Then, two distributions were created from the weighted sum of estimated raw scores for each panelist, one for the attribute profiles judged proficient and one for the attribute profiles judged not proficient. The cut score was established as the point that most distinguished between the distributions. The average of these cut scores across panelists was used as the final cut score. An example of these two distributions for a panelist is shown in Figure 2.1.

In addition to the DP methods, the authors conducted a second standard setting with the same number of panelists, the same test items and item parameters of the TIMSS and two rounds of standard setting, but utilized the Modified Angoff Method within an IRT framework. The authors compared the cut scores produced by the Modified Angoff Method and the DP method

Figure 2.1

Weighted Distribution and Cut Score for a Single Panelist



Note. The arrow represents the cut score location for one panelist. Reproduced from Skaggs et al. (2016, p. 450).

and found the cut points produced by the two methods to be within two raw score points. They also found that the standard deviation of panelists' cut scores was higher for the Angoff standard setting group.

Condensed Mastery Profile Method for Setting Standards

The second approach for standard setting extant in the DCM literature was conducted by Clark, Nash, Karvonen, and Kingston (2017) and is called the Condensed Mastery Profile Method. The Condensed Mastery Method was used to set the cut points for a large-scale assessment system call the Dynamic Learning Maps (DLM) Alternate Assessment. This assessment was developed by a consortium of 15 states and two Bureau of Indian Education tribal schools for federal accountability testing of students with significant cognitive disabilities. This method is substantially different than the approach taken by Skaggs et al. (2016) and is described below.

In comparison to the simulated DCM used in Skaggs et al. (2016), the DLM assessment was designed from the start to utilize a DCM. First, the consortium specified attributes that represented the knowledge, skills, and abilities (KSAs) students are expected learn. Then, these attributes were organized into a series of maps that model the expected hierarchies and learning progressions of those KSAs. The attributes, which DLM call essential elements (EE) but we will continue to call attributes, were measured in testlets of three to eight items each at one of five linkage levels. The five linkage levels were related attributes ordered in complexity, with three precursor levels called *initial precursor*, *distal precursor*, and *proximal precursor*, a *target* level that represented grade-level performance, and a *successor* level beyond the target. Teachers can assign DLM testlets to students on demand. Students that show mastery of a testlet at a higher

linkage level (i.e., a more complex attribute within the five attribute linkage levels) are assumed to have mastered all lower levels.

To conduct the CMP standard setting, the authors first worked with the consortium to develop policy PLDs for three levels of proficiency. After that, they selected profiles for the standard setting by identifying the three most common profiles of student mastery at each total mastery level that differed on at least three attributes, where total mastery level is the total number of attributes mastered. For example, the three most prevalent profiles were selected for all students that had mastered 20 attributes, but which had at least three attributes different among the 20 mastered. The 20 attributes mastered in this example is the total mastery level for these three profiles. The authors asserted that this provided some variation and prevented the profiles from being overly similar. Once the profiles were selected, the authors created PDFs that visually displayed profiles with rows of different attributes and columns representing the five linkage levels, which were the ordering of the attributes by complexity. These tables were then shaded to show mastery of the attributes, with a light green shading representing the mastery of an attribute and no shading to represent nonmastery. Figure 2.2 (on the following page) presents the visual display of the profiles presented to panelists.

Having selected and developed the profiles, the authors then created panels of experts to classify the profiles. They found content experts and educators with expertise working with students with significant cognitive disabilities. They created a total of fourteen panels with four to eight members each and recommended that future practitioners use between four to six panelists. Before the panelists made any ratings, Clark et al. provided them with both in-advance

Figure 2.2

Example Profile Used in the Condensed Mastery Profile Standard Setting

Essential		Level mastery				
Area	Element	Initial Precursor	Distal Precursor	Proximal Precursor	Target	Successor
ELA.C1.1	ELA.RL.4.1	Identify familiar people, objects, places, or events	Identify character actions in a familiar story	Identify character actions	Recount events in a story using details	Recount the key details of a story
ELA.C1.2	ELA.RL.4.2	Identify familiar people, objects, places, or events	Identify major events in a familiar story	Identify a character's actions and corresponding consequences	Identify the theme of a familiar story	Identify the specific theme of a story
ELA.C1.1	ELA.RL.4.3	Understand object names	Identify concrete details in a familiar story	Identify characters, setting, and major events	Describe characters in a narrative	Describe characters, setting, and events
ELA.C1.2	ELA.RL.4.4	Understand object names	Identify the meaning of words	Identify words or phrases to complete a literal sentence	Identify the meaning of an unambiguous word	Identify multiple meanings of a word
ELA.C1.1	ELA.RL.4.5	Identify familiar people, objects, places, or events	Name or identify objects in pictures	Identify the beginning, middle, and end of a familiar story	Identify story characteristics	Identify story elements that change
ELA.C1.2	ELA.RL.4.6	Understand object names	Identify character actions in a familiar story	Identify character actions	Identify the narrator of a story	Identify narrator point of view
ELA.C1.1	ELA.RI.4.1	Understand object names	Name or identify objects in pictures	Identify concrete details in an informational text	Identify explicit details in informational texts	Identify words related to explicit information
ELA.C1.1	ELA.RI.4.2	Understand object names	Name or identify objects in pictures	Identify concrete details in informational texts	Identify the overall topic of a familiar text	Identify topic-related words in an informational text
ELA.C1.1	ELA.RI.4.3	Understand object names	Use category knowledge to draw conclusions	Identify concrete details in an informational text	Identify concrete details related to people, events, or ideas	Compare key details
ELA.C1.2	ELA.RI.4.4	Understand object names	Identify the meaning of words	Identify words or phrases to complete a literal sentence	Identify the meaning of an unambiguous word	Identify the multiple meanings of a word

Note. Reprinted from Clark et al. (2017).

training before the standard setting and in-person training at the start of the standard setting event. Panelists were given a quiz after the in-advance training to communicate their level of comfort with the training material and convey any remaining questions. A practice standard setting was held at the end of the in-person training in which panelists viewed several profiles and practiced providing ratings of profiles for three proficiency levels. The panelists discussed how they made their profile ratings and facilitators practiced collecting the ratings.

The main portion of the standard setting consisted of two rounds of range finding and two rounds of pinpointing. Each panel was responsible for classifying profiles from assessments from three grade levels in either English language arts or mathematics. With between 30 and 50 attributes per assessment, the DLM assessments had a considerable number of attributes. The range-finding was meant to identify the general location of divisions between performance levels. During this round, the panelists were provided with folders with three student profiles in increments of five total mastery levels (e.g., one folder with three profiles at 5, 10, 15, etc. total mastery levels). Independently, panelists reviewed profiles and categorized them into performance levels. Panelists could ask facilitators to project sample items for each attribute onto a display. After all ratings had been made facilitators recorded the ratings of each panelists by a show of hands into a predeveloped Excel file. This data was presented to the panelist who then discussed the ratings and major points of disagreement. These ratings were also used in analysis after the standard setting to see if panelists increased consensus between rounds. After the discussion, panelists were given an opportunity to revise their categorizations during a second round of ratings. After the second round of ratings was conducted, the facilitators used the Excel file to find the point of maximum uncertainty between performance levels using a built-in logistic regression function. The regression function found the value for which being classified

into two contiguous performance levels was .5 since this represents the point of maximum disagreement. The total number of attributes mastered associated with this point was used to select a smaller set of profiles for the pinpointing in the next stage. The authors selected the three most prevalent profiles at that level, and at the three total mastery levels below and above this level. This resulted in 21 profiles for each of the two cuts between the three performance levels. For example, if the point of maximum uncertainty was 40 total attributes mastered, then three profiles were selected for 37, 38, 39, 40, 41, 42, and 43 total attributes mastered.

During the pinpointing, panelists independently categorized each profile on a rating sheet. Ratings were shared with the group by a show of hands and recorded by the facilitator in a workbook. As with the range finding, a discussion and second round of rating ensued. Once these ratings were made and collected, the logistic regression function produced a value of the cut point. The facilitators adjusted the cut points based on the proportion of students in each categorization between grades, essentially smoothing the data so that the percent of students in a performance level did not differ substantially from year to year.

The authors analyzed the results of the standard setting by considering the convergence of ratings, standard errors of pinpointing ratings, and surveys from panelists. The convergences of ratings, or the building of consensus of judges' ratings, were analyzed by creating box-and-whisker plots, with each successive round represented by a different box and whisker plot on the same chart. Visual inspections showed a clear convergence of the plots with the boxes and whiskers becoming less spread as each round of rating and discussion brought panelists closer together towards consensus. The authors analyzed the standard error of the panelists' pinpointing ratings for each of the four performance levels on all of the 40 assessments ($n = 160$), with a

range of .08 to 1.25 and a median of .20; these values indicate that the multiple rounds of ratings produced a small overall amount of variability in the final ratings.

The panelists completed evaluations of the standard setting that asked about the training they received, the process of setting standards, and their beliefs about the appropriateness of the resulting cut points. The responses indicated the panelists believed the training prepared them for the process and that the process was sound. Further, when asked if they would adjust the cut points, 95% stated they would not make any adjustments.

A final aspect to note is that Clark et al. (2017) did not use impact data during the main portion of the standard setting. The documentation of the standard setting on the Dynamic Learning Maps website reports that impact data was provided to the technical advisory committee (TAC) after the cut points were established through the standard setting procedure described above (Clark, 2015). The TAC made slight alterations to the cut points to make the number of students with each designation of proficiency more proportional across grades.

Standard Setting Approach Proposed in Georgia's IADA Application

The state of Georgia has applied and been approved by the Federal Department of Education to pilot an assessment system that produces diagnostic results under the Innovative Assessment Demonstration Authority (IADA). The Putnam Consortium uses a diagnostic classroom assessment system called Navvy and created by Navvy Education, LLC. As a part of their application for the IADA, their system provided a description of how they will conduct standard settings. A description of the approach is described below.

Navvy is an on-demand assessment system, where teachers can administer assessments of specific standards at any point throughout the semester. These assessments use a CDM to

provide mastery diagnostics on individual state standards and deliver results immediately. Some students will take both the innovative assessments and the Georgia Milestones. In the application to for the IADA (Georgia Department of Education, 2018), the Putnam Consortium described multiple methods for making the single summative determination required by federal policy. The first was using the intact attribute profiles to make the determination, rather than consolidating the profiles with multiple mastery designations into a single numerical result. To do this, they intend to use non-parametric clustering methods to map the profiles onto the nearest Georgia Milestones Achievement level. Then, PLDs will be written based on the mastered standards for each achievement level. This is arguably a more valid way to develop the PLDs since this method would be in line with the underlying design of the diagnostic assessment system and would be based on empirical data produced by the assessment system rather than general arbitrary descriptions of performance developed by policymakers.

The next approach proposed by Navvy was to use summative scores, such as percentages of attributes mastered or weighted percentages of attributes mastered, and logistic regression, similar to what was done by Clark et al. (2017), except replacing the panel judgements with results from the Georgia Milestones tests. This is likely an easier approach in that the data might more closely map onto a logistic regression model than a clustering approach and is a generally simpler model that might be more easily communicated to stakeholders.

The final approach proposed by the Putnam consortium is for students that do not take both the innovative assessments and the legacy state assessments. In these cases, teachers will be asked to place students into achievement levels. Then, based on these teacher designations, the results from Navvy assessments, and Milestones achievement from students in affiliate districts that take both tests, student performance level will be established using one of the two methods

described above. Then, the achievement level distributions for each grade and subject will be smoothed to ensure the resulting distributions of this method produce meaningful interpretations across grades and subjects (Georgia Department of Education, 2018, p. 95). This approach seems suitable during a pilot program but may require greater validation for use at scale. It is possible that teachers are overly dependent on their experience with their own students and therefore produce judgmental bias or error.

Implications of Previous Methods for a DCM Standard Setting Approach

Several aspects of the methods discussed here hold implications for a DCM standard setting approach. First, the limitations of the Angoff, Modified Angoff, and Bookmark Methods serve as guides to the development of a standard setting approach for DCMs. It is evident that the cognitive load demanded from panelists can present a challenge during a standard setting. Whether imagining an individual barely proficient student or choosing the percent of one hundred such students, several of the studies described above point out the difficulty for panelists in making predictions of student behavior. A DCM standard setting method needs to consider the cognitive demands it places on panelists and attempt to mitigate anything that would make the approach overly burdensome.

An additional aspect that should be considered while developing an approach for DCM standard setting in light of previous methods is the concept of the response probability or mastery threshold. For the IRT standard setting methods, the response probability was important because the panelists were expected to conceptualize the RP for a student on each item. In a DCM standard setting, panelists view profiles of students instead of test item. The response probability is essentially replaced by the mastery threshold, or the required probability necessary to be designated as a master of an attribute. DCMs produce probabilities of mastery for each

attribute that range from 0 to 1. Responses closer to 0 represent greater certainty of nonmastery, responses closer to 1 represent greater certainty of mastery and responses closer to .5 represent less certainty of categorizations. Different assessment systems can place the threshold at different levels in order to manage the chance of a false categorization. In the Diagnostic Profiles method used by Skaggs et al. (2016), the authors used a mastery threshold of .5. In the Condensed Mastery Profile Method used by Clark et al. (2017), the assessment system, based on discussions with their Technical Action Committee (TAC), decided to use a threshold of .8 in order to reduce the chance that a student was incorrectly designated as a mastery. The profiles presented to panelists were selected based on their prevalence in the data and changing the mastery threshold is likely to change the prevalence of profiles. It is therefore important to consider the level of the mastery threshold and how it could impact the resulting selected profiles or cut point.

Research Questions

The remainder of this chapter discusses the research questions of this study. These questions are based on the review of the literature presented above and with an eye towards solving the research problem presented in the previous chapter.

Question 1: How Does the Set of Profiles Presented to Panelists During a DCM Standard Setting Impact the Resulting Cut Point?

DCMs produce attributes where each attribute A can take on either a 1 or 0 for mastery or nonmastery status, respectfully. That means that there are 2^A possible attribute profiles that represent all possible combinations of mastery or nonmastery among the attributes for a DCM. For example, for a DCM with 14 attributes there are 2^{14} or 16,384 possible attribute profiles. With even a relatively small number of attributes, it is impossible to show all profiles to the

panelists. The process of selecting profiles and the number of profiles presented to panelists becomes an important input in the standard setting process with potentially substantial effects on the resulting cut points. It is not clear, for example, if panelists will make the same cut points if they see two profiles at each attribute mastery level compared to if they see three profiles at each mastery level. This is one aspect of the DCM standard setting that will be explored in this study.

Question 2: How Does the Variability of the Profiles Presented to Panelist During a DCM Standard Setting Impact the Resulting Cut Point?

In addition to the process of selecting the profile, the previous methods of DCM standard setting have all shown the same set of profiles to the panelists. Expanding the set of profiles so that panelists within the same standard setting see different sets of attribute profiles (i.e., adding variability) could result in a more informed standard setting. For example, providing variability in the set of profiles may result in greater discussion during standard setting sessions or differences in the resulting cut point because panelists will see profiles throughout the distribution of possible profiles. This has played out in the IRT literature where more discussion was found to increase the consistency of expert judgements of student performance (Clauser et al., 2009), but has yet to be studied in a DCM context.

Question 3: How Does Presenting Only Part of the Total Attribute Mastery Range to Panelists Impact the Resulting Cut Point Location?

A third question that arises for a DCM standard setting approach is if panelists consistently choose the same cut point location if they are presented with only part of the attribute mastery range. Clauser et al. (2017) examined the consistency of judgements made by panelists in IRT settings by providing ordered item booklets of varying difficulty to judges during several bookmark standard settings. They found that content experts were not able to

make consistent judgements between the booklets of varying difficulty suggesting a lack of internal consistency. This question is aimed at a similar question in a DCM context.

In previous DCM standard setting approaches the panelists were given example profiles at different levels of total attributes mastered. For example, Clark et al. (2017) presented panelists with profiles of student that mastered no attributes all the way to profiles of students that had all attributes mastered and asked panelists to make cuts along that range. There has yet to be experiments examining if the range impacts the cut points. If the panelists are able to establish an objective cut point, the portion of the range of total attributes mastered presented to panelists should not impact the resulting cut point. It may be that seeing only attribute profiles from part of the range of all possible total attributes mastered will influence how panelists set their cut points. To put another way, the domain of attributes may impact the resulting cut raising important questions about the objectivity of the standard setting process.

Question 4: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Panelists'; (a) Confidence in Their Ratings, (b) Agreement with the Resulting Cut Point, and (c) Perception of the Fairness and Validity of the Cut Point Setting Process?

Manipulating the variables described in the previous questions could result in a change in the perceptions and agreement of panelists. At the end of the standard setting, panelists will be asked to complete a survey that asks about their confidence in the rating they made, their agreement with the resulting cut points that the panel made, and about their perception of the standard setting including the fairness and validity of the process. It is important for the validity of the standard setting process that those involved in it are confident in the rating they and the group make and that they do not identify a lack of fairness or validity in the process. The

responses from the exit survey will serve to determine their confidence, agreement, and perceptions of the standard setting across different conditions.

Question 5: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Observed Participation of Panelists?

A crucial aspect of the standard setting is the participation of panelists. To this end, the facilitator will record the level of participation and the amount of discussion among the panelists during the standard setting. The facilitator will take notes after the standard setting of their observations and impressions. In addition, these recordings will take place by recording the standard setting and then analyzing the resulting records. These notes and recordings will provide evidence for how the conditions of the standard setting impact the participation of the panelists.

Chapter 3. Study Design

This chapter discusses the study design and the steps that will be taken to answer the research questions. It begins here with a brief, general outline of the process and then details various aspects of the study. It describes the establishment of the research conditions, how panelists will be identified and recruited, how the profiles panelists will see will be created, the steps for conducting the standard setting, and the analysis of results.

For this study, the general process will begin by identifying five sets of five experts to serve as panelists in a standard setting for fourth grade mathematics. All panelists will complete a recruitment survey, receive an asynchronous in-advance training, complete a post in-advance training survey, participate in one of five remote but synchronous standard settings, and complete a post standard setting survey. Appendix F presents a facilitator guide that will be used to guide the standard setting. Each set of five panelists will receive a different set of profiles which will create five experimental groups for the study. The conditions for each experimental group are described in the subsequent section. Institutional review board approval for this study has been granted. Institutional review board (IRB) approval has been received and the IRB approval letter is presented in Appendix I.

Establishing Conditioning

The first step to answering the research question is to establish the experimental conditions. There will be five conditions to answer the research questions. The empirical data used in the study is described followed by a description of each condition considering the research questions.

The selection process of the profiles seen by each condition of this study will rely on empirical data from an operational diagnostic assessment system. The empirical data will be used to determine the prevalence of attribute profiles. In several of the study conditions, the prevalence of the attribute profiles is used to determine which profiles are presented to panelists.

The empirical data will be from Navvy, which generates attribute profiles from a CDM. Attributes in this system are based on standards that are aligned to the Georgia Standards of Excellence. Each student that uses Navvy receives a competency or noncompetency diagnosis on each standard. Navvy provided the base rates, or the percent of students with a specific student profile, for the 4th grade mathematics assessments. These base rates will be used to determine the prevalence of profiles that will be used in the study conditions. Next, the study conditions with reference to the study questions are described.

Conditions for Questions 1: How does the Set of Profiles Presented to Panelists During a DCM Standard Setting Impact the Resulting Cut Point?

Condition 1 represents the control and is similar to what was used by Clark et al. (2017) in the Condensed Mastery Profile Standard Setting Method. In this condition, each panelist will see the three most prevalent profiles at each mastery-level with different mastery statuses on at least three attributes where possible. The profiles from the operational diagnostic assessment system will be ranked by their prevalence in the data at each level of total attributes mastery. The most prevalent profiles will be selected at each total mastery level as the first profile. The second profile will be the next most prevalent profile that has at least three attributes different than the first and the third will be the next most prevalent profile that has at least three different from the first and second. In cases where multiple profiles have an equal prevalence and at least three

attributes different from the previous selection, a random number generator built into R will be used to select from amongst the potential profiles. In cases where there is not a profile that is at least three attributes different, a situation Clark et al. (2017) reported, the profile with the next greatest number of attributes different will be selected.

The standard setting will be broken into two parts, range finding and pinpointing. During the range finding, panelists in Condition 1 will be presented with the three most prevalent profiles, with at least three attributes mastery statuses different, at each total attribute mastery level in increments of three (3, 6, 9, etc.). This approach is comparable to the presentation of profiles in increments of five used by Clark and colleagues. A logistic regression function is used with the rating of panelists from the range finding to generate a location for the pinpointing along the domain of the total attributes mastered. During pinpointing, the panelists of Condition 1 will see the three most prevalent profiles, with at least three attribute statuses different, at the cut point generated during the range finding, as well as the three most prevalent profiles with at least three attribute mastery statuses different at each of the two mastery levels above and the three profiles at each of the two mastery levels below the cut point generated during range finding. For example, if the logistic regression model gave six as the cut point, during the pinpointing stage this condition will see three profiles at four, five, six, seven, and eight total attributes mastered.

Condition 2 is meant to determine if there is a significant difference in the cut point location when the number of profiles is reduced at each step described in Condition 1. Whereas Condition 1 panelists see three profiles at each total level of attributes mastered during the range finding and pinpointing stages, Condition 2 panelists will see only two profiles at each level. By comparing the resulting cut scores of this condition with that of Condition 1, a general effect of presenting panelists with different numbers of profiles can be determined. It would be beneficial

to know if reducing the number of profiles seen by panelists can maintain consistent results because viewing fewer profiles would presumably create a less demanding cognitive load and time of panelists.

Conditions for Question 2: How does the Variability of the Profiles Presented to Panelist during a DCM Standard Setting Impact the Resulting Cut Point?

In the previous standard setting methods for DCMs described in the chapter 2 literature review (Clark et al., 2017; Skaggs et al., 2016), all panelists saw the same set of profiles. While this may make the process more uniform, it limits the total number of profiles viewed by all panelists. Condition 3 explores the scenario where panelists at the same standard setting see different sets of profiles. Specifically, this method uses a choose without replacement approach. As with Condition 1, the profiles are ranked by their prevalence with the top 15 profiles chosen as the pool. Each panelist will receive three different profiles from this pool chosen so that each set of three balances the rankings of the profiles by prevalence. See Table 3.1 for the balancing of the profiles prevalence rankings for each panelists. In this condition, each panelist will see a different set of three profiles at each level of total attributes mastered. This condition will, in comparison to the previous conditions, explore how variability of the profiles impacts the standard setting process and results. It is expected that variations of the profiles may create greater discussion and impact the consistency of the ratings.

Table 3.1*Prevalence of Each Profile Viewed by Panelists in Condition 2*

Panelist	First profile	Second Profile	Third Profile	Summed Rank
1 st Panelist	1	8	15	24
2 nd Panelist	2	9	13	24
3 rd Panelist	3	10	11	24
4 th Panelist	4	6	14	24
5 th Panelist	5	7	12	24
Total	15	40	65	120

Note. This table presents the ranking of the profiles viewed by each panelist. The total ranking of the profiles is equal for each panelist.

Conditions for Question 3: How Does Presenting Only Part of the Total Attribute Mastery

Range to Panelists Impact the Resulting Cut Point Location?

The last two conditions are closely related. For Condition 4 and Condition 5, the panelists will see the same profiles as those seen by Condition 1, however, they will see a reduced range. Condition 4 will see only profiles in the bottom approximate two-thirds of all levels of total attributes mastered and will be asked to only set the lowest two cut points. With 28 attributes for fourth grade mathematics, they would see only the attribute profiles that have a total attribute mastery of 0 to 20. Condition 5 will see only profiles in the top two-thirds of total level of attributes mastered and be asked to set the top two cut points. For fourth grade mathematics they will view only profiles with between 8 and 28 attributes mastered. These conditions provide a means to examine the internal consistency of the judgements made by the panelists. Clauser et al. (2017) examined if the cut points of judges differed when given ordered item booklets from the same test but with items with different average item difficulties. They found that the judges' cut scores were substantially impacted by the difficulty of the test. The cut points of these conditions

will be compared to the cut points set by Condition 1 and to each other. If the range of the total attributes mastered does not impact the standard setting, we would expect to see similar cut points across conditions.

Panelists Selection Criteria

To be included in the study, participants will need to meet several criteria. First, all participants will need to be familiar with the 4th grade mathematics standards for Georgia. Second, they will need to have direct experience working with the 4th grade student population that are required to take federal accountability tests. The educators do not need to currently be working with 4th grade students or work only with 4th grade students but should have at least one year of experience working directly with this population in mathematics. A final criterion is that the participants will need to be available during the days and times of the standard settings.

Panelists Recruitment

Recruitment efforts will target teachers and school mathematics specialists in the state of Georgia. A recruitment letter and or email (see Appendix A) will be sent to schools and teachers in Georgia. Emphasis will be placed on individuals with 4th grade mathematics expertise and experience working with 4th grade mathematics students. All candidates will be asked to complete a Qualtrics survey (see Appendix B). Survey items will include general demographic information and questions about experience and expertise. Panelists will also be asked to confirm their ability to assure confidentiality, to complete an in-advance training, and to their availability for the remote standard setting meetings.

Five panelists will be recruited for the five condition groups in addition to four backup panelists who will complete the training and be asked to join the standard settings if other panelists back out. Having back up panelists was a recommendation from a professional standard setting expert. The Qualtrics surveys will be evaluated to select the panelists who maximize the amount of expertise and that include a range of experience. Panels will be formed based on availability to attend one of the five sessions. If multiple panelists are able to attend the same or multiple sessions, they will be divided so that teachers from the same school and then district attend different sessions, to reduce school or district effects. If there is further freedom to alter group membership, teachers' years of experience will be used to determine the panelists' assigned condition so that each panel's average years of experience is maximized and roughly equivalent. The aim of using availability, school, district and years of experience is to create roughly equivalent groups. This nonrandom sampling approach is called purposive sampling and, according to Etikan and colleagues (2016), is appropriate when a researcher needs to select individuals that are experts in a subject and availability to participate is an important factor. All panelists will receive an honorarium of 125 dollars for their participation in the study.

Creating the Profile Cards and Panelists Packets

The profile cards were created using a table in Microsoft PowerPoint. Each cell of the table represents one of the attributes which is also one of the Georgia State Standards of Excellence. There are 29 standards for 4th grade mathematics and after some adjustment, the standards fit best in an eight by four table. The formatting of the table is a 13 inch high by 17 inch wide page. An example profile card is presented in Appendix J. The standards were color coded with grey representing the nonmastery and green representing mastery. The state standards

are also grouped into five domains (Operations and Algebraic Thinking, Number and Operations in Base Ten, Number and Operations—Fractions, Measurement and Data, and Geometry).

Because educators would likely be familiar with these domains and after discussion with the pilot participants (discussed below) and measurement experts, the proportion of each domain that was mastered was listed in the top left box of the profile.

In advance of the standard setting, each panelist will receive a set of profiles by mail. The packet will include a set of training profiles for the practice standard setting (described in the Training Panelists: Introduction and Before Standard Setting Training), a set of profiles for the range finding session, and a set of profiles for the pinpointing session. Between the range finding and pinpointing phase, a list of profiles will be provided to panelist and they will be asked to remove extra profiles from the packet. This is necessary because the starting point of the pinpointing session is dependent on the results of the range finding. Each profile card will be clearly labeled at the top of the page.

Training Panelists: In-Advance Training

Once selected, the panelists will participate in an asynchronous training in advance of meeting, an in-person introduction and before standard setting training, a standard setting range finding, and a standard setting pinpoint (see Appendix E for a schedule of panelists activities). The in-advance training will last approximately 30 minutes. The presentation used during the training is presented in Appendix H. The training will provide panelists with an overview of their roles and expectations, how to interpret results of the operational diagnostic assessment system including student profiles, the standard setting purpose and process, a description of the student population, the standards, and the profiles, and the performance level descriptors (PLDs). The PLDs are the

general (not the grade or subject specific) PLDs used by the state of Georgia for mathematics and are presented in Table 3.2. The training will be delivered in the form of a video and a set of accompanying resources to include a pdf of the PowerPoint used during the training, a copy of the relevant standards, a copy of the PLDs, and sample student profiles. At the end of the training, participants will be guided to the post in-advance training survey that will ask how a copy of the relevant standards, a copy of the PLDs, and sample student profiles. At the end of the training, participants will be guided to the post in-advance training survey that will ask how familiar they feel with the panelists' expectations, how confident they feel about the standard setting procedures and process, and their understanding of the student population, standards, profiles, and PLDs (See Appendix C). The survey will also provide a space for the panelists to ask questions that will be discussed during the in-person training session as needed.

Table 3.2

Georgia Performance Level Descriptors

Beginning Learner	Developing Learner	Proficient Learner	Distinguished Learner
Beginning Learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students need substantial academic support to be prepared for the next grade level or course and to be on track for <i>college and career readiness</i> .	Developing Learners demonstrate partial proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students need additional academic support to ensure success in the next grade level or course and to be on track for <i>college and career readiness</i> .	Proficient Learners demonstrate proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students are prepared for the next grade level or course and are on track for <i>college and career readiness</i> .	Distinguished Learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students are well prepared for the next grade level or course and are well prepared for <i>college and career readiness</i> .

Training Panelists: Introduction and Before Standard Setting Training

The standard setting will begin with introductions and a before standard setting training that will review the key elements of the standard setting and last approximately 30 minutes. The review will include a discussion of any lingering questions from the post in-advance training survey, the panelists expectations, purpose of the standard setting, confidentiality expectations, and a discussion of the standards, profiles, and PLDs. This will lead to a practice standard setting. During the practice standard setting, panelists will utilize the set of practice profiles they received via mail (described above) and a link delivered to them via email to a Qualtrics survey that will collect profile categorizations. They will work independently to categorize the set of practice profiles into each of the performance levels and mark their judgements in the survey. This will be followed with a discussion of how the panelists made their judgements and a discussion of profiles that received the greatest variability in categorization.

Standard Setting: Range Finding

After the practice standard setting, panelist will receive a new set of profiles depending on the condition group they are in and be asked to take 45 minutes to mark their ratings in a new Qualtrics survey. During the range finding, panelists in Condition 1 will see three profiles from each level of total attributes mastered in increments of three (3, 6, 9, 12, 15, 18, 21, 24, and 27 total attributes mastered). This is similar to the range finding of Clark et al. (2017) that saw three profiles at mastery levels in increments of five (0, 5, 10, 15, etc.). Panelists in Condition 2 will see two profiles at each level of total attributes mastered in increments of three. The panelists of Condition 3 will see three profiles at each mastery level in increments of three, with the profiles selected from a pool of the fifteen more prevalent profiles and selected for each panelist based on

the rank sum of the profile, as described in table 3.1. The panel of Condition 4 will see the three most prevalent profiles at each total attribute mastery level between 0 and 21 total attributes mastered in increments of three (3, 6, 9, 12, 15, 18, 21 total attributes mastered). Condition 5 which will see the three most prevalent attribute profiles between 9 and 28 total attributes mastered in increments of three (9, 12, 15, 18, 21, 24, 27 total attributes mastered). Once all panelists have completed their ratings, the panelists will report their ratings to the facilitator. A summary of the ratings will be presented on the screen and a 30-minute discussion will ensue. Panelists will discuss the ratings they made and be guided through discussion by the facilitator. Panelists will have an opportunity to change their ratings before submitting the Qualtrics survey.

After all surveys have been received, the data will be analyzed using a logistic regression model. This will be accomplished by first dummy coding the profile ratings based on whether they span each of the proficiency level distinctions. That is, if a profile is rated as Beginning Learner, it will be assigned a 0 for the Beginning/Developing Learner proficiency cut because it does not span the proficiency level distinction. If a profile is designated as Developing Learner, it is assigned a 1 on the Beginning/Developing Learner proficiency cut. If a profile is rated as Proficient Learner, it will receive a 1 for both the Beginning/Developing Learner proficiency cut and the Developing/Proficient Learner proficiency cut, but still receive a 0 for the Proficient/Distinguished Learner proficiency cut. This will be done for each rating of each profile by each panelist. An example of this coding is presented in table 3.3. Each profile rating is given its own row and the panelists are repeated for each rating. The total number of attributes mastered is given its own column and is, obviously, the same for each profile ID. In the table, panelists 1 rated Profile A14, with 14 total attributes mastered, as proficient and it is therefore coded with a 1 in the Beginning/Developing Learner and Developing/Proficient Learner

Table 3.3*Example Coding of Panelists Ratings*

Profile ID	Panelist	Total Attributes Mastered	Beginning/ Developing Learner	Developing/ Proficient Learner	Proficient/ Distinguished Learner
A0	1	0	0	0	0
...
A14	1	14	1	1	0
A14	2	14	1	1	0
A14	3	14	1	0	0
A14	4	14	1	0	0
A14	5	14	1	1	0
A15	1	15	1	1	1
...
A28	5	28	1	1	1

columns, but a zero in the last column. That same panelists rated the profile A15 as

Distinguished, and therefore the every column receives a 1 since all cuts are spanned.

Using a single ordinal regression model for all cut points was explored as an alternative to multiple individual logistic regressions models for each cut point. This approach was not used because testing found that the results for the range finding rounds would be the same, regardless of which approach was used. However, during the pinpointing rounds (discussed below) the panelists will not be given the opportunity to rate the subset of profiles they see into all categories. During the pinpointing round panelists will instead only be allowed to rate the profiles they see into one of two adjacent categories for profiles of five total attributes mastery levels centered on the results of the range finding. This means that the levels would not be completely ordinal in the sense that a panelist could not categorize the profile into all levels of proficiency. Furthermore, the potentially large gaps in the selected ranges of the cut points also meant that the predictor variable would not be completely continuous. As such, a single ordinal regression model would not be appropriate.

Once the data is entered, an excel file will automatically code the data and a logistic regression model will be fit for each of the three proficiency level cuts (Beginning/Developing, Developing/Proficient, Proficient/Distinguished). The dependent variable will be the vector of dummy coded profiles (the three right columns of table 3.3) and the independent variable will be the total mastery level. The equation for the logistic regression model is of the form

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where $p(x)$ is the probability that a profile with a total attribute mastery level of x is categorized in the higher of the two adjacent proficiency categories, β_0 and β_1 are the intercept and slope of the log-odds of the function. Using the ratings of the panelists and solving for β_0 and β_1 provides a logistic regression function of the probability a profile with a total attribute mastery level of x is in the higher of the two adjacent proficiency categories. Using this function, the point where the probability of being classified is equal to .5 is chosen as the starting mastery level, rounded to the nearest whole number, of the pinpointing stage discussed in the following section. This location was chosen because it represents the point of maximum disagreement among the panelists and the clearest distinction between the categories. This value will be called the initial proficiency cut point.

Standard Setting: Pinpointing

The pinpointing portions of the standard setting will occur once each of the initial proficiency cut points is calculated. The profiles from two total mastery levels below the initial proficiency cut point, the initial proficiency cut point, and two total mastery levels above the initial proficiency cut point will be made into a list and distributed to the panelists. Panelists will be asked to remove all unused profiles from their packets before beginning the ratings. This is to

ensure that they do not rate the wrong profiles. A new rating survey will be distributed to panelists via email. The set of profiles for Conditions 1 and Condition 3 will include 45 profiles (three at each of the five levels surrounding three cut points). The set of profiles for Condition 2 will include 30 profiles (two at each of the five levels surrounding three cut points). The set of profiles for Condition 4 and Condition 5 will include 30 profiles (three at each of the five levels surrounding two cut points).

The panelists, who will have been on a 30-minute break, will come back together and conduct another round of independent ratings. The ratings will be reported to the facilitator who will present the groups' ratings to the screen and a discussion will occur highlighting those places with the greatest disagreement. Finally, a second round of independent ratings will occur where panelists can change their responses.

Independent Evaluation and Exit Survey

After all panelists have completed their responses for the pinpointing, a second logistic regression model will be run. Final cut points will be reported to panelists. Then, panelists will complete an exit survey and independent evaluation (see Appendix D). This evaluation will ask the degree to which panelists believe the procedure used to find the cut points was suitable and if the cut points were appropriate. They will be asked which of the specific cut points, if any, they do not agree with, what they believe a more appropriate value would be, and why they do not agree.

Post Standard Setting Facilitator Notes

After the standard setting, the facilitator will answer a set of questions about the standard setting (see Appendix G). The questions ask about the degree of participation of the panelists, the level of discussion and interaction among the panelists, and how the panelists came to

consensus (or not) around ratings. These notes will be used to qualitatively assess the standard setting process with variation of the conditions.

Pilot Study

A pilot study of the methods was conducted to hone the steps and approach taken during the standard setting. The participants for the panel included three graduate students in an educational psychometrics PhD program and a university faculty professor that specializes in educational psychometrics. One of the graduate students had extensive experience with item writing and DCM test development. The faculty professor was a specialist in DCMs.

The pilot helped in several ways. First, the pilot provided recommendations about panelist training, including preparing questions for the discussion, highlighting that the order of the profiles does not necessarily relate to the categorization. This resulted in an adjustment of the training material. The pilot participants also suggested adding several questions and a back button to the survey given after the in-advance training survey and the post standard setting exit survey. Next, the pilot participants provided feedback on the visual displays presented to panelists. Participants recommended improving the student profile cards by presenting the fractions of the portion correct of the content domains of each profile and clearer labels, both of which were added.

The pilot participants also made several recommendations related to the standard setting process, including providing the panelists with regular reminders to look at the achievement level descriptors throughout the standard setting process. They advised changing the submission of profile recommendations in Qualtrics by adding a separate page that asked participants to wait for discussion before submitting final results. In addition, the pilot helped to work out what questions to ask panelists during the practice standard setting and standard setting rounds. These

questions related to how participants think about different levels of proficiency, approaches to asking participants about their proficiency designations to spark discussion, and questioning panelist on the strategies they used to make decisions.

Data

The data of this study will take multiple forms and is presented here in a table 3.4 as a summary on the following page

Table 3.4

Table of Data of this Study

Data Name	Summary of Data
Post In-Advance Training Survey	Ask how familiar panelists feel about the expectations, how confident they feel about the standard setting procedures and process, and their understanding of the student population, standards, profiles, and PLDs (See Appendix C).
First Ratings	The initial independent ratings of profiles made during the range finding.
Second Ratings	The independent ratings of profiles made during the range finding stage after seeing other panelists ratings.
Third Ratings	The initial independent ratings of profiles made during the pinpointing.
Fourth Ratings	The independent ratings of profiles made during the pinpointing stage after seeing other panelists ratings and discussion.
Exit Survey	The evaluation will ask the degree to which panelists believe the procedure used to find the cut points were was suitable and if the cut points were appropriate. They will be asked which of the specific cut points, if any, they do not agree with and why (See Appendix E).
Post Standard Setting Facilitator Notes	Ratings of the facilitator about the degree of participation, discussion, and interaction among the panelists, and how the panelists came to consensus around ratings. (See Appendix G).

Instruments

This study uses several instruments. The following table 3.5 presents the instruments and their location.

Table 3.5

Table of Instruments of this Study

Instrument	Location
Panelists Recruitment Survey	Appendix B
Post In-Advance Training Survey	Appendix C
Range Finding Categorization Survey	Qualtrics
Pinpointing Categorization Survey	Qualtrics
Standard Setting Exit Survey	Appendix D
Post Standard Setting Facilitator Notes	Appendix G

Analyses

We will analyze the data collected throughout the standard setting in several ways. We will examine the data descriptively by checking the convergence of condition groups ratings between rounds, computing and comparing the standard error of the final cut points, conducting a nonparametric, longitudinal, factorial analysis of the panelists ratings and following that with planned contrasts, comparing the panelists' post-standard setting feedback and agreement with final cut scores, and by analyzing the notes taken by the facilitator. These approaches are discussed below in greater detail in reference to each question.

Analysis for Question 1: How Does the Set of Profiles Presented to Panelists During a DCM Standard Setting Impact the Resulting Cut Point?

The analysis for the first question will occur by visually comparing convergence plots, computing the standard errors, and conducting a nonparametric, longitudinal factorial analysis. Convergence plots are vertical box plots that show the quartiles of the ratings of the judges, with

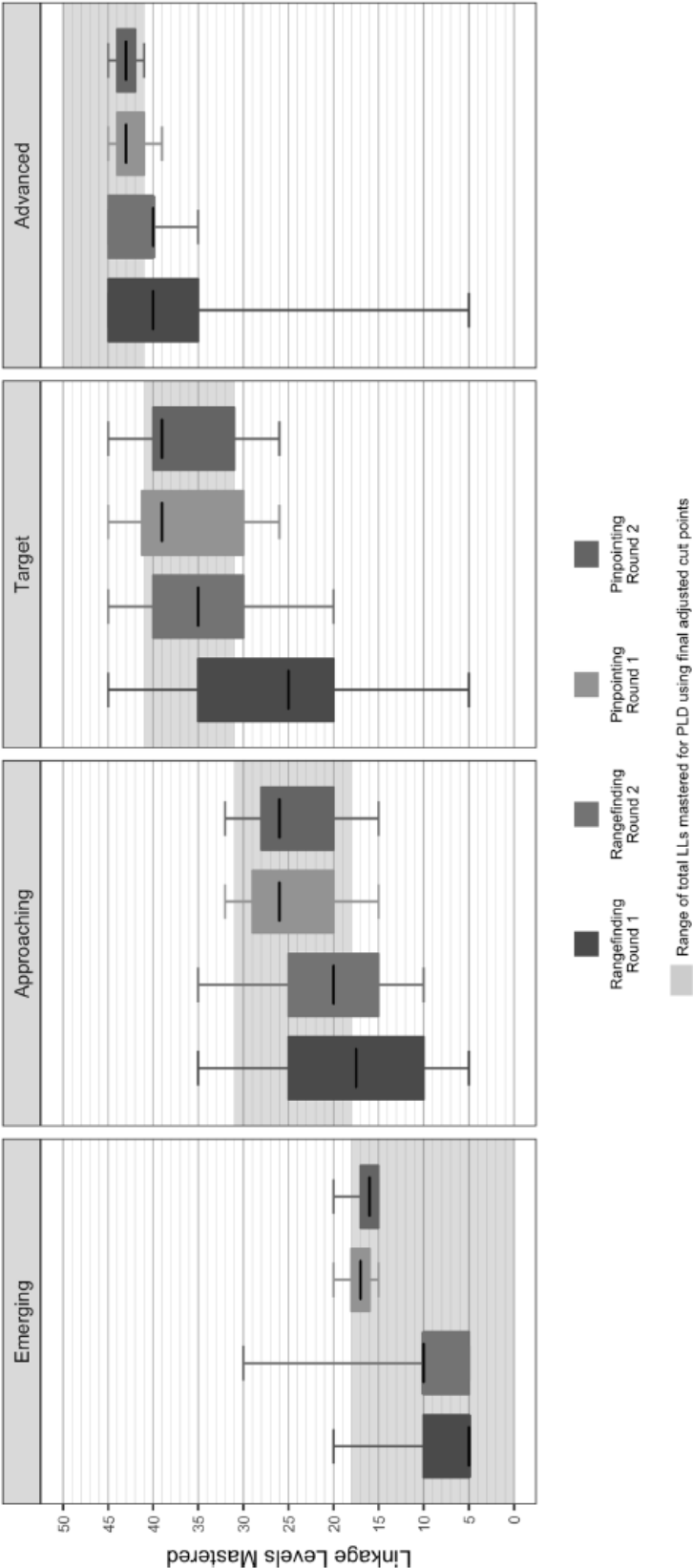
the total mastery level along the y-axis. The plots are grouped by the proficiency level for which the judges are making their ratings and set in a series to show the change of ratings between rounds (the first categorization during range finding before discussion, second categorization during range finding after discussion, third categorization during pinpointing before discussion, and fourth categorization during pinpointing after discussion). An example set of convergence plots is reprinted from Clark et al. (2017) in Figure 3.1 on the following page. In our case, however, we will have a second set of plots comparing the results of different panels. The plots are expected to show that the panelists come to increasing levels of convergence in their ratings between rounds. Comparisons of the convergence plots for Condition 1 and Condition 2 will be the first step in examining if the set of profiles presented to panelists impacts the resulting cut score. We expect that the panelists in the first condition will come to consensus faster because they will view more profiles and therefore have more discussion.

After visually comparing the convergence plots, we will compute the standard error of the final cut points. This will be computed by dividing the standard deviation of the frequencies of panelists' final cut points by the square root of the number of total ratings. These ratings will be presented in a table with the conditions of the other questions and can be compared across conditions. If the standard error values are generally small, we can interpret the results to mean there is a small amount of variability in the final ratings.

After visually comparing the convergence plots and computing the standard error of the cut scores, we will conducting a nonparametric, longitudinal factorial analysis with follow up planned contrasts (Brunner & Puri, 2001; Konietzschke et al., 2010). We will use nonparametric

Figure 3.1

Example Convergence Plots



Note. Reprinted from Clark et al. (2017).

analysis because of the limited sample size ($n = 5$ panelists per condition) and because model assumptions, such as sphericity or an observable distribution may not be met. The lack of sphericity is expected because we believe that the ratings will converge during the rounds of ratings and therefore there will not be the same variance over time. The analysis is longitudinal because the panels will provide ratings of profiles multiple times ($t = 4$). Finally, the analysis is factorial because there are both condition groups ($a = 5$) and multiple cut points ($b = 3$ cut points, i.e., beginning/developing learner, developing/proficient learner, and proficient/distinguished learner). We will use the nparLD package in R to conduct this analysis (Noguchi et al., 2012).

The analysis will begin with an omnibus test of significance and use an ANOVA-type statistic (ATS; Brunner & Puri, 2001) and an alpha (α) value of .05 (level of significance) to determine if there are significant effects of time, treatment, or cut score. If a significant result is found, we will conduct multiple contrasts test procedure (MCTP) that allows for inference about pairwise comparisons between samples and controls for the family wise error rate (Konietschke et al., 2010). To answer the first question, a contrast will be created to compare Condition 1 (the control) and Condition 2. The contrasts matrix of this study is

$$C = \begin{pmatrix} c'_1 \\ c'_2 \\ c'_3 \\ c'_4 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & .5 & .5 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

where c'_1 represents the comparison of the first and second condition, and the remaining contrasts are discussed below in relation to the relevant question. By comparing the first and second condition, we will determine if there is an effect of presenting panelists with different sets of profiles.

In addition to reporting the results of the above test and planned contrast, we will report the relative marginal effects for each of the factors. In nonparametric analysis, there are no parameters in the model and so relative marginal effects are used to describe an effect (Brunner & Puri, 2001). The control group will act as the baseline and the effect of being in the other conditions will be reported for each of the treatment effects and any possible interaction effect. Effect sizes will be reported in a table that shows the relative marginal effect for each factor and condition.

An important aspect of the nonparametric ANOVA is the power of the test, or the ability to find a result if there is one, that is avoiding a type II error. Brunner et al. (1997) conducted simulation studies that demonstrated that the power of the marginal effects analysis is high and comparable to ANOVA for normally distributed data. Furthermore, Shah and Madden (2004) found that approximating the ATS with the F statistic is reasonable and works well with moderate or small sample sizes. Therefore, we conducted a power analysis using the G*power software (Faul et al., 2009).

During the power analysis, we utilized the F test statistics for ANOVA with repeated measures with within and between interactions. Alpha was set to $\alpha = .05$ and beta was set to $\beta = .2$. Because no data is available to establish expected effect size, we used small, medium, and large effects of $f = .10$, $f = .25$, and $f = .40$, respectively, recommended by the G*power manual (Faul et al., 2009). When conducting power analysis for repeated measure designs, the correlation among the repeated measures is an input parameter. By utilizing the standard error of the cut scores and sample sizes reported by Clark et al. (2015), we chose .86 as the correlation among repeated measures.

The results of the power analysis indicate that for small effect sizes we would need 65 total participants, for a medium effect size we would need 15 total participants, and with a large effect size only 10 total participants. Furthermore, increasing the correlation among repeated measures reduced the required sample size in all cases. These results indicate that if no significant effect is found and the effect size is small, we may fail to find a significant result if there is one with five participants per group. In this regard, feasibility limits our study, however, if the effect size is medium to large, we should have plenty of power to find a significant result.

Analysis for Question 2: How Does the Variability of the Profiles Presented to Panelist During a DCM Standard Setting Impact the Resulting Cut Point?

As with the first question, convergence plots of Condition 1 and Condition 3 will be compared to see if adding variability to the attribute profiles viewed by panelists in Condition 3 affects the convergence of judge's ratings. In addition, the standard error of the final cut points of this condition will be included in the table of standard errors and can be used to compare the resulting variability of the Condition 1 and Condition 3. Furthermore, c'_2 compares the first and third condition and will test if there is a statistically significant difference between the groups. Comparing these condition groups will determine if adding variability to the profiles impacts the resulting cut points. We predict that adding variability will lead to more discussion for the third condition, and as a result it will produce better convergence of the ratings.

Analysis for Question 3: How Does Presenting Only Part of the Total Attribute Mastery Range to Panelists Impact the Resulting Cut Point Location?

As with the previous two questions, convergence plots will be visually compared between Condition 1, Condition 4 and Condition 5 to see if there are differences in the resulting cut points when panelists are presented with only part of the total attribute mastery range. In addition, the standard error of the final cut points of Condition 4 and Condition 4 will be included in the table of standard errors and can be used to compare the resulting variability of the conditions. The nonparametric factorial analysis described above can handle unbalanced designs (Konietschke et al., 2010). This design is unbalanced because there will be more ratings by the panelists of Condition 1, who see the whole scale and create three categorizations, than the panelists of either Condition 4 or Condition 5 that see only a part of the scale and make only two categorizations. Contrast c'_3 compares the results of Condition 1 with a combination of Condition 4 and Condition 5. Contrast c'_4 compares the results of Condition 4 and Condition 5. Is it expected that Condition 4 will have significantly lower cut points and Condition 5 will have significantly higher cut points because of the differences in the range the panelists see.

Analysis for Question 4: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Panelists'; (a) Confidence in Their Ratings, (b) Agreement with the Resulting Cut Point, and (c) Perception of the Fairness and Validity of the Cut Point Setting Process?

To answer the fourth question, we will compare the results of the standard setting exit survey (Appendix D). The exist survey includes the judges' confidence in their ratings, their agreement with the resulting cut points, and their beliefs about the fairness and validity of the

standard setting process across conditions groups. We will compare the feedback provided by the panelists to see if any conditions result in different perspectives of the standard setting. The confidence in their ratings and their perception of the fairness and validity of the standard setting process are reported on a five-point Likert scale. We will sum up the responses to these questions for each response and compare the results across conditions. For the panelists' agreement with rating, we will count and compare the number of panelists that believe the responses are too high, appropriate, or too low for each group. In addition, after stating they believe a cut point is too high or too low, the panelists will receive a follow up question asking what they believe would be a more appropriate value. We will sum up the absolute value of the difference between the condition groups final cut point and the value the panelists state is more appropriate. Using these summations, we will compare the responses of panelists to see if there are differences in the panelists' opinions of the standard setting process and results across conditions.

Analysis for Question 5: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Observed Participation of Panelists?

To answer the fifth question, we will utilize the post standard setting facilitator notes (Appendix G) to analyze the degree of participation of the panelists, the level of discussion and interaction among the panelists, and how the panelists came to consensus (or not) around ratings. This analysis will consider if any differences were noticed between condition groups. Further, we will report on what aspects of the standard setting conditions and differences between the conditions led to variations in the group participation, interaction, and level of consensus.

Chapter 4. Results

This chapter provides the results of the standard setting study. It includes a discussion of the results from panelists recruitment efforts, process of selecting panelists, the data collection process, the resulting recommended cut scores, the changes of ratings between rounds, the convergence of the ratings over rounds, the standard deviations and standard errors of the cut points, a discussion of the nonparametric factorial analysis, and the results of the panelists exit survey ratings.

Panelists Recruitment

The dates of the standard setting were selected based on an analysis of district academic calendars as published on the state department of education website. Dates with the most overlap of vacation days were selected for the standard settings in the hope that those dates would provide for the greatest number of available educator panelists. Panelists were recruited by sending emails to every district superintendent in the state of Georgia, every principal in the state of Georgia of a school that included fourth grade, and a list of 6023 teachers and math specialists from 742 elementary schools in the state of Georgia, representing roughly 57% of all school districts. 120 respondents took the recruitment survey.

Panelists Selection

30 panelists were selected as participants, 25 as the actual panelists and five backup panelists using a purposive sampling method (Etikan et al., 2016). To select panelists, they were

first organized into groups based on which dates they selected in the recruitment survey as having availability for participation. Some potential participants were in multiple groups because they indicated availability for multiple days. They were then ordered based on their years of experience. Educators with the most experience were contacted first. The next contacted educator was the one that indicated availability for the next standard setting date (which was also the next condition group) and with the next most years of experience that did not work at the same school as the previous educator in that group. In this way, participants were added to each condition group, cycling through the groups, maximizing the years of experience and minimizing any school or district effects in that group. Some potential participants responded to initial notification of selection stating that their availability had changed, in which case the next potential participant on the list was contacted utilizing the same selection process as described above. After the interest survey and after the initial contact and confirmation of participant desire to join the study, participants were provided with a link to the training video (<https://www.youtube.com/watch?v=jpCChVdhveI>) and the post in-advance training survey (Appendix C). Across the entire study, 14 school districts were represented. Only two educators in the study were from the same school and they were in different condition groups. No more than two educators from the same school district were in the same condition group.

Table 4.1 provides the mean, median, minimum, maximum, and standard deviation of the years of experience of the panelists in education in general, in mathematics, and in fourth grade mathematics specifically, and the panelists' ratings of their own familiarity with the 4th grade mathematics standards. Overall, the participants had a great deal of experience and familiarity, with 12.2 average years of experience in education, 10.8 average years of experience in mathematics, 6.8 average years of experience in 4th grade mathematics specifically, and an

Table 4.1*Panelists' Experience and Familiarity with Standards by Condition Group*

	Mean	Median	Minimum	Maximum	Standard Deviation
Years of experience as educator overall					
Condition 1	9.2	9	5	16	4.32
Condition 2	10.2	9	7	16	3.42
Condition 3	17.8	19	6	28	8.90
Condition 4	10.4	11	7	13	2.41
Condition 5	13.4	11	7	20	5.77
All Participants	12.2	10	5	28	5.92
Years of experience with mathematics					
Condition 1	7.6	8	5	10	2.07
Condition 2	8.6	9	6	12	2.61
Condition 3	17.2	18	6	28	8.56
Condition 4	9.6	11	6	12	2.51
Condition 5	11.0	10	6	20	5.57
All Participants	10.8	10	5	28	5.68
Years of experience with 4 th grade					
Condition 1	5.4	5	3	8	2.07
Condition 2	5.0	5	3	8	1.87
Condition 3	10.0	4	3	24	9.25
Condition 4	5.2	6	2	8	2.28
Condition 5	8.2	7	5	15	3.96
All Participants	6.8	5	2	24	4.82
Familiarity with 4 th Mathematics (0 unfamiliar to 10 very familiar)					
Condition 1	9.4	10	7	10	1.34
Condition 2	9.0	9	8	10	1.00
Condition 3	9.2	10	7	10	1.30
Condition 4	9.0	9	8	10	0.71
Condition 5	9.0	9	9	9	0.00
All Participants	9.1	9	7	10	0.94

**Note.* This table includes only those panelists that ultimately participated in the study.

average rating of 9.1 out of 10 points for familiarity with the 4th grade mathematics standards. Generally, the condition groups were well balanced. Condition 3 looks slightly different than the other groups because one of its participants had 28 years of experience, considerably more than any other participant. Excluding this participant, Condition 3 looks much more similar to the others.

Post In-advance Training Survey

After watching the in-advance training video, participants completed a survey about how familiar they felt with the expectations of panelists, how confident they felt about the standard setting procedures and process, and their understanding of the student population, standards, profiles, and PLDs (see Appendix C). The results of this survey are presented in Table 4.2. It shows that for 260 out of 312 ratings (83%) of the statements, participants expressed either a "Good" or "Excellent" rating. There were some statements with lower ratings, including "Characteristics of Students who take DCM assessments," "How testlets measure the intended content," "How testlets are made accessible to students and teachers," "What a student is expected to do during a DCM Assessments," and "How DCM results are reported." Of the 120 ratings of these 5 questions, 38 (32%) were either "Poor" or "Fair," they account for all ratings of "Poor" in the entire survey, and they account for 73% of the ratings that were not "Good" or "Excellent" in the entire survey. These aspects were targeted and addressed during the in-person training at the start of each standard setting to shore up any misunderstandings.

Table 4.2*Results of Post In-Advanced Training Survey*

	Poor		Fair		Good		Excellent		Total
	Count	%	Count	%	Count	%	Count	%	Count
Characteristics of students who take DCM Assessments	2	8.33	4	16.67	7	29.17	11	45.83	24
How prepared you feel to take part in the standard setting process	0	0.00	2	8.33	14	58.33	8	33.33	24
The purpose of standard setting	0	0.00	1	4.17	14	58.33	9	37.50	24
The 4th grade Mathematics Georgia Excellence Standards	0	0.00	0	0.00	9	37.50	15	62.50	24
The 4th grade Mathematics Achievement Level Descriptors	0	0.00	1	4.17	10	41.67	13	54.17	24
How to interpret a student profile	0	0.00	4	16.67	15	62.50	5	20.83	24
The expectations for maintaining security of information	0	0.00	0	0.00	5	20.83	19	79.17	24
How testlets measure the intended content	1	4.17	7	29.17	9	37.50	7	29.17	24
How testlets are made accessible to students and teachers	2	8.33	9	37.50	9	37.50	4	16.67	24
What a student is expected to do during a DCM Assessments	3	12.50	6	25.00	11	45.83	4	16.67	24
How DCM results are reported	2	8.33	4	16.67	12	50.00	6	25.00	24
Discussing achievement level descriptors with other panelists	0	0.00	2	8.33	8	33.33	14	58.33	24
Discussing profiles with others	0	0.00	2	8.33	9	37.50	13	54.17	24
Total	10	3.21	42	13.50	132	42.31	128	41.03	312

Data Collection

Data collection occurred over one week in February of 2023. One participant failed to show for Condition 5 of the study and a backup panelist also failed to maintain availability, resulting in that condition group having only four panelists. Otherwise, data collection occurred as intended and described in the previous chapter. After introductions and a brief review of the training material (Appendix H), a practice standard setting was conducted with panelists rating profiles and engaging in a practice discussion.

Panelists were then instructed to take out of the manila envelopes that had been mailed to them the set of profiles for the range finding stage. The profiles mailed to panelists were selected using the process described in Chapter 3 under the Creating the Profile Cards and Panelists Packets section. These packets were printed on high grade 11x17 inch paper with laser inkjet printers to make the print clear and easily legible. The profiles for each panelist were divided into two sets, one for the range finding and one for the pinpointing stages. Panelist could use the first set of packets readily upon receipt to review the profiles for the first two rounds of ratings, and then by unclipping the easy-to-use ring and integrating the packets together could rate the profiles for the last two rounds of ratings.

Panelists then received a link to an online form to enter and conduct the first round of ratings. As described in the previous chapter and displayed concisely in Table 4.4, the set of profiles presented to panelists differed by the number of profiles panelists viewed at each level of total attributes mastered, whether the panelists saw the same or different profiles within the same condition group, and the range of the profiles panelists saw across the levels of total attributes mastered. Data was collected and presented to panelists. Appendix M shows an example of the excel spreadsheet used to display the results of the initial round of range finding to panelists for

Table 4.4*Set of Profiles Presented to Panelists During Range Finding Rounds*

Condition	Set of Profiles Viewed by Panelists	Profiles Per Level	Range of Total Attributes Mastered	Total Profiles Viewed	Cut Points Set
1	Same 3 profiles at every total attribute mastery level	3	3. 6. 9, 12. 15. 18, 21, 24, 27	45	3
2	Same 2 profiles at every total attribute mastery level	2	3. 6. 9, 12. 15. 18, 21, 24, 27	30	3
3	Different sets of profiles for each panelist at every total attribute mastery level	3	3. 6. 9, 12. 15. 18, 21, 24, 27	45	3
4	Same 3 profiles at every total attribute mastery level	3	3. 6. 9, 12. 15. 18	30	Lower 2
5	Same 3 profiles at every total attribute mastery level	3	12. 15. 18, 21, 24, 27	30	Upper 2

discussion. The spreadsheet showed a table with the name of each profile on each row, the name of each panelist in the column header, and the rating of each panelist in the intersection. These ratings were in numeric form for easy readability, with 1 for Beginning Learner, 2 for Developing Learner, 3 for Proficient Learner and 4 for Distinguished Learner. The final column of the table included a calculation of the variance of each row as a simple measure of disagreement of the panelists. The facilitator chose profiles that represented the greatest degree of variance in panelists' ratings and that included ratings that spanned the range of total attributes mastered and cut points. For example, the facilitator first chose a profile that had high disagreement among panelists at the middle range of the total attributes mastered scale, that was rated as both Developing Learner and Proficient Learner. Once discussion of that profile concluded, the facilitator chose a profile that was rated as both a Beginning Learner and Developing Learner, usually at the lower end of the total attributes mastered level. This process

was repeated with the Proficient and Distinguished Learner ratings, and then rotating back through the profiles that represented the highest degree of disagreement as time permitted. During discussion, the facilitator presented the profile on the screen and asked panelists to explain why they rated the profile as they did and promoted discussion among the panelists. After the allotted time for discussion, panelists conducted a second independent rating for the range finding using the same set of profiles.

The facilitator took the results of the second range finding round and conducted a logistic regression model for each cut point. The results of the logistic regression are presented in Table 4.5. The values served as the middle of the range (after rounding to the nearest whole number) of the set of five consecutive profiles rated by panelists in the pinpointing rounds. For example, for Condition 1 the logistic regression for the Beginning to Developing Learner cut point resulted in a value of 8.15. The set of profiles rated by panelists during the pinpointing round in this condition as either Beginning or Developing was two levels of total attributes mastered above the cut point, the level of total attributes mastered at the cut point, and two levels of total attributes mastered below the resulting cut point. In this case, it was the set of profiles at six, seven, eight, nine, and ten total attributes mastered.

Table 4.5

Result of the Logistic Regression Functions Conducted After the Second Round of Range Finding

Condition	Beginning / Developing Learner	Developing / Proficient Learner	Proficient / Distinguished Learner
1	8.15	17.31	23.75
2	9.00	15.29	22.49
3	8.30	16.90	24.76
4	6.30	15.34	NA
5	NA	14.48	22.78

Table 4.6 provides information on the sets of profiles presented to panelists during the pinpointing rounds. In the column labeled “Range of Total Attributes Mastered,” the ranges of attributes presented to each condition group are presented. The values in Table 4.5 are reflected in this column as the center of the ranges in that column. These ranges are based on the results of the logistic regression models developed for each cut point of the range finding stage. Panelists again conducted independent ratings and the data was collected and presented to panelists in the same spreadsheet described above for the range finding rounds and presented in Appendix L.

The results of running a logistic regression for each cut point at this round are presented below in Table 4.7. Specifically, the rows labeled “Pinpointing” under the column labeled “Stage” and “1” under the column labeled “Round” correspond the resulting cut points of this round.

Table 4.6

Set of Profiles Presented to Panelists During Pinpointing Rounds

Condition	Set of Profiles Viewed by Panelists	Profiles Per Level	Range of Total Attributes Mastered	Total Profiles Viewed	Cut Points Set
1	Same 3 profiles at every total attributes mastery level	3	6-10, 15-19, 22-26	45	3
2	Same 2 profiles at every total attributes mastery level	2	7-11, 13-17, 20-24	30	3
3	Different sets of profiles for each panelist at every total attributes mastery level	3	6-10, 15-19, 23-27	45	3
4	Same 3 profiles at every total attributes mastery level	3	5-9, 14-18	30	Lower 2
5	Same 3 profiles at every total attributes mastery level	3	12-16, 21-25	30	Upper 2

After collecting the data, the facilitator guided participants through a discussion of those profiles with the greatest disagreement at each cut point, in a similar approach to that taken during the discussion in between the range founding rounds. Panelists were then given a chance to alter their responses a final time. The data was collected a final time, results were run through the logistic regression function script in R, and final cut points were presented to panelists. The results of the logistic regression functions are presented and discussed in the next section.

Standard Setting Recommended Cuts Points

Cut point recommendations for each round, each condition group, and each rated proficiency level were estimated by conducting a logistic regression on the ratings from each condition group, cut point, and round. Table 4.7 shows the final values for each cut point for each condition group and each round of rating. The actual final cut points were rounded to the nearest whole number. These results are also summarized visually in Figure 4.1, which provides a stacked bar chart for each cut point with a separate bar for each round grouped by condition.

Several distinctions are evident between the conditions and several patterns are evident across the rounds. First, Conditions 4 and Condition 5 look different than Conditions 1, 2, or 3. These two condition groups did not rate the Distinguished Learner achievement level and Beginning Learner achievement level, respectfully, as indicated by the lined bar for the regions they did not rate. Furthermore, it is evident that the beginning learner cut points for Condition are lower than for any other group by between two and three total attributes mastered. Another difference between the groups is that Condition 3 has the least amount of change across rounds

Table 4.7*Final Cut Points Produced by Logistic Regression Functions by Condition Group and Round*

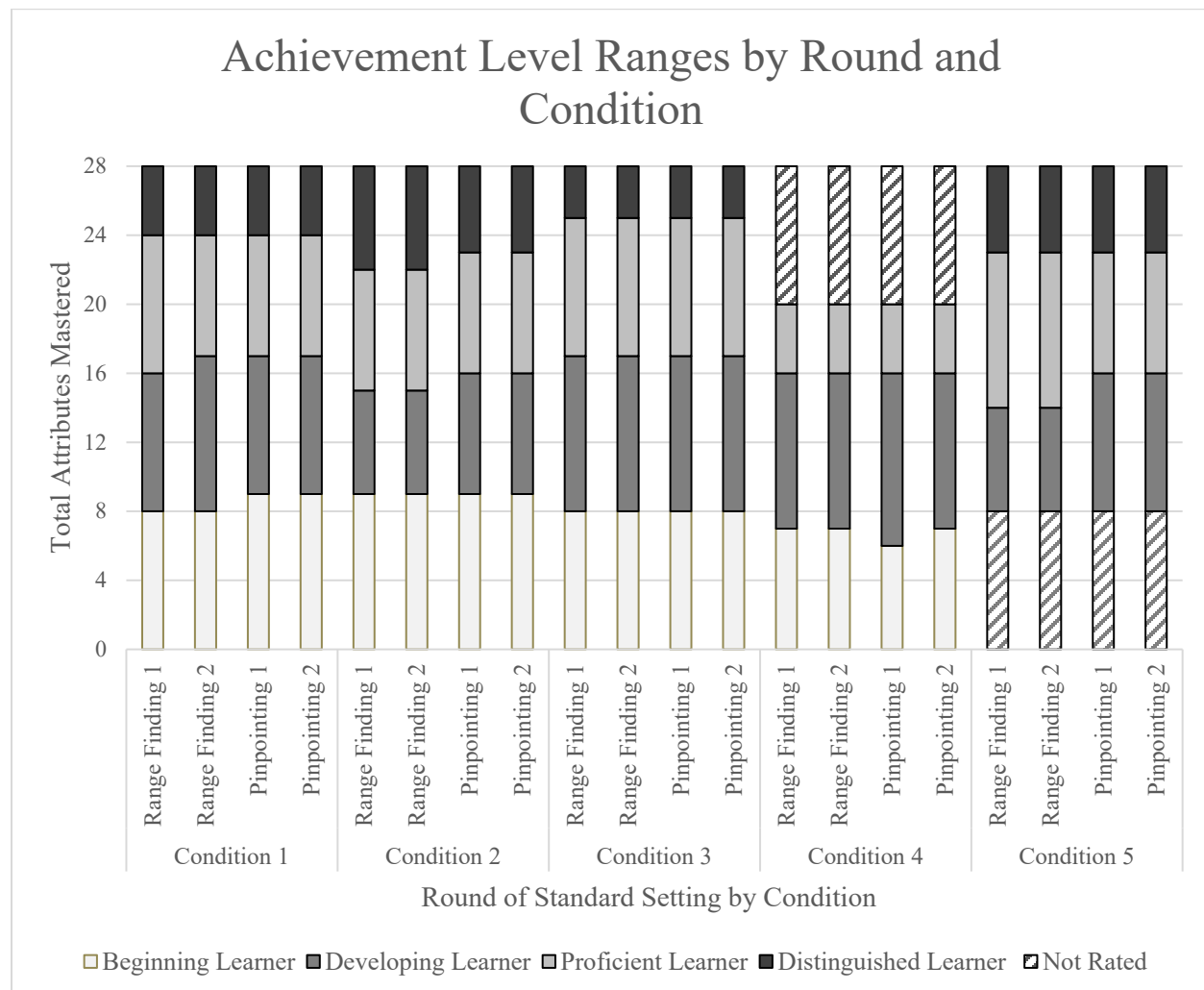
Condition	Stage	Round	Beginning / Developing Learner	Developing / Proficient Learner	Proficient / Distinguished Learner
1	Range Finding	1	7.91	16.30	23.52
		2	8.15	17.31	23.75
	Pinpointing	1	8.78	16.97	23.62
		2	8.88	17.26	23.57
2	Range Finding	1	8.94	15.12	22.49
		2	9.00	15.29	22.49
	Pinpointing	1	9.42	16.06	23.13
		2	9.42	16.06	22.53
3	Range Finding	1	8.30	17.10	24.76
		2	8.30	16.90	24.76
	Pinpointing	1	7.80	17.03	24.98
		2	7.71	17.03	24.98
4	Range Finding	1	6.54	15.96	NA
		2	6.30	15.34	NA
	Pinpointing	1	6.17	15.86	NA
		2	6.59	15.61	NA
5	Range Finding	1	NA	13.96	22.56
		2	NA	14.48	22.78
	Pinpointing	1	NA	15.85	22.77
		2	NA	15.83	22.58

Note. Actual final ratings were rounded to nearest whole number.

for all three cut points. Whereas Condition 1 and Condition have changes for the middle and upper cut points across rounds, and Condition 4 has changes for the lower cut points across rounds, and Condition 5 has changes for the middle cut point across rounds, Condition 3 is the only group that has no change of cut point ratings across rounds. These cut points will be discussed further in light of each question in the next chapter.

Figure 4.1

Final Achievement Level Ranges by Round and Condition Group



Note. Condition 4 and 5 did not rate ratings at the upper and lower ranges, respectfully, as indicated by the “Not Rated” label.

Change and Convergence of Rating

The design of the standard setting was meant to build consensus among the panelists over the rounds of ratings. In between the first and second rounds of both the range finding and pinpointing stages, discussions were held on the profiles that provided the greatest degree of

disagreement. We expected to see consensus building manifest in changes in panelists ratings and a narrowing of the range and degree of disagreement of their ratings.

The interrater agreement of profiles for panelists in each condition and in each round was calculated using Fleiss' Kappa. This measure is appropriate for measuring the interrater agreement of categorical data with more than two raters (Fleiss, 1971). The results are presented in Table 4.8. When interpreted Fleiss' Kappa, values less than 0.20 indicate slight agreement, values between 0.21 and 0.40 indicate fair agreement, values between 0.41 and 0.60 indicate moderate agreement, values between 0.61 and 0.80 indicate substantial agreement, and values between 0.81 and 1.00 indicate near perfect agreement (McHugh, 2012, p. 279). Several interesting aspects of the results are worth noting. First, between every round where discussion took place, the change in the Kappa value was positive. This indicates that in all cases the interrater agreement increased. Next, it's worth noting that with Kappa values 0.43 and 0.38 for the range finding and pinpointing round and changes of only +.01 for each round, Condition 3 had the lowest degree of interrater agreement and the least amount of change between rounds. Furthermore, though the interrater agreement is the lowest, it is still considered moderate for the

Table 4.8

Interrater Agreement of Profile Categorizations for Each Condition and Round

Condition	1 st Range Finding	2 nd Range Finding	Change	1 st Pinpointing	2 nd Pinpointing	Change
1	0.55	0.80	+0.24	0.56	0.73	+0.17
2	0.87	0.88	+0.01	0.49	0.56	+0.07
3	0.43	0.44	+0.01	0.37	0.38	+0.01
4	0.49	0.55	+0.06	0.40	0.41	+0.01
5	0.46	0.53	+0.07	0.36	0.46	+0.10

range finding round and fair for the pinpointing round. This is important because these panelists did not see the same sets of profiles. The values indicate that the panelist rated profiles with similar numbers of attributes mastered ‘moderately’ and ‘fairly’ similarly, even if the profiles did not have the same attributes mastered. Further, the fact that there was little change in the value indicates that panelists resisted changing their categorizations because they did not see a common set of profiles. Next, it is worth noting that with a Kappa of 0.87 and 0.88, the range finding stages of Condition 2 had the highest interrater agreement among any of the stages or rounds. It seems that the reduced set of profiles viewed by this group of panelists induced the most agreement. Further exploration of this was conducted by comparing the rating of the profiles in Condition 1 to the profiles in Condition 2. The profiles of Condition 2 were the same set of profiles, just reduced from the three most prevalent profiles at each mastery level to the two most prevalent profiles at each mastery level. The comparison of the common set of profiles found in the sets of profiles between Condition 1 and Condition 2 found that there was 92% agreement in the ratings and a cross condition Kappa rating of 0.86. This indicates that the two most common profiles at each level produce much more interrater agreement and that introducing the third profile at each total attribute mastery level produces a good deal more disagreement in the ratings.

The change of ratings over rounds is displayed in Table 4.9, which shows the number and percentage of rating changes disaggregated by panelists and by condition groups over the rounds of discussion. Several aspects are worth noting. First, there was a lack of trend in the categorization changes with a relatively similar number of decreases and increases within each condition group. Next, in line with the interrater agreement discussed above and the box-and-whisker convergence plots discussed below, Condition 3 has nearly no changes between rounds.

Table 4.9*Condition and Panelist Rating Changes by Round*

	RF	Decreases	Increases	PP	Decreases	Increase
Condition 1	17 (12.59)	9 (6.67)	8 (5.93)	19 (8.44)	8 (3.56)	11 (4.89)
Panelist 1	2 (7.41)	2 (7.41)	-	3 (6.67)	2 (4.44)	1 (2.22)
Panelist 2	0 (0)	-	-	3 (6.67)	-	3 (6.67)
Panelist 3	2 (7.41)	2 (7.41)	-	3 (6.67)	1 (2.22)	2 (4.44)
Panelist 4	6 (22.22)	2 (7.41)	4 (14.81)	6 (13.33)	3 (6.67)	3 (6.67)
Panelist 5	7 (25.93)	3 (11.11)	4 (14.81)	4 (8.89)	2 (4.44)	2 (4.44)
Condition 2	5 (5.56)	1 (1.11)	4 (4.44)	10 (6.67)	9 (6.00)	1 (0.67)
Panelist 6	0 (0)	-	-	0 (0)	-	-
Panelist 7	0 (0)	-	-	2 (11.11)	1 (5.56)	1 (5.56)
Panelist 8	3 (16.67)	-	3 (16.67)	3 (16.67)	3 (16.67)	-
Panelist 9	1 (5.56)	-	1 (5.56)	2 (11.11)	2 (11.11)	-
Panelist 10	1 (5.56)	1 (5.56)	-	3 (16.67)	3 (16.67)	-
Condition 3	1 (0.74)	1 (0.74)	0 (0)	1 (0.44)	1 (0.44)	0 (0)
Panelist 11	0 (0)	-	-	0 (0)	-	-
Panelist 12	0 (0)	-	-	0 (0)	-	-
Panelist 13	0 (0)	-	-	0 (0)	-	-
Panelist 14	0 (0)	-	-	0 (0)	-	-
Panelist 15	1 (3.7)	1 (3.7)	-	1 (2.22)	1 (2.22)	-
Condition 4	5 (4.76)	1 (0.95)	4 (3.81)	7 (4.67)	6 (4.00)	1 (0.67)
Panelist 16	0 (0)	-	-	0 (0)	-	-
Panelist 17	2 (9.52)	-	2 (9.52)	3 (10)	3 (10.00)	-
Panelist 18	1 (4.76)	1 (4.76)	-	0 (0)	-	-
Panelist 19	2 (9.52)	-	2 (9.52)	4 (13.33)	3 (10.00)	1 (3.33)
Panelist 20	0 (0)	-	-	0 (0)	-	-
Condition 5	9 (8.57)	4 (3.81)	5 (4.76)	4 (2.67)	1 (0.67)	3 (2.00)
Panelist 21	5 (23.81)	3 (14.29)	2 (9.52)	1 (3.33)	-	1 (3.33)
Panelist 22	0 (0)	-	-	0 (0)	-	-
Panelist 23	2 (9.52)	1 (4.76)	1 (4.76)	0 (0)	-	-
Panelist 24	2 (9.52)	-	2 (9.52)	3 (10)	1 (3.33)	2 (6.67)

Note. “-” indicates no change. RF = Range Finding Rounds, PP = Pinpointing Rounds. Values in parentheses represent the percentage of ratings changed for that condition group or that panelists in that round. The total number of profile ratings was different between groups and is discussed in the pinpointing and range finding sections.

In addition, panelists in many conditions or in some stages were either revisionists who changed several categorizations or non-revisionists who made no alternations. Eight of the 24 panelists made no revisions in either round. Consider the 48 rating sessions described below (two for each panelist), there were 19 sessions where panelists made no revisions. Meanwhile, the five panelists with the most categorization changes (Panelists 4, 5, 8, 19, and 21) accounted for 52% of all changes. Finally, while some conditions had a higher count of changes, apart from Condition 3, the percentage of rating changes was generally similar. This is important to consider because the conditions saw different numbers of profiles and made different numbers of profile ratings.

The convergence of the panelists' ratings are also displayed in box-and-whisker plots in Appendix N to visually represent the ratings over rounds. For each round of the standard setting, these plots show the minimum, first quartiles, median, third quartile, and maximum of the ratings in terms of the total attribute mastered, grouped by proficiency level. Careful inspection of these plots shows that there was some consensus building between rounds for most of the conditions. The rounds of interest are after the first round of range finding and pinpointing, when panelists discussed profiles and then adjusted their ratings during the second rounds of these stages. The changes between the second round range finding and first round pinpointing is influenced by the results of the logistic regression functions applied to the data and the shift from profiles at every three total attributes mastered to profiles at five consecutive levels of total attributes mastered. For most conditions the narrowing of ratings for the middle cut points of Developing Learner and Proficient Learner is clearer than the narrowing for either the Beginning Learner or Distinguished Learner, likely because the latter two are bounded. That is, the Beginning Learner proficiency cut point is bounded by the all-nonmastery profile, and the Distinguished Learner

proficiency cut point is bounded by the all-mastery profile. This meant that a certain level of consensus was built into the design of the study, as evident by the smaller range of the ratings of these proficiency level ratings. In terms of convergence, some conditions appear to have more than others. Notably, Condition 1 appears to have the most and Condition 2 appears to have the second most, whereas the latter two conditions show less convergence. Condition 3 shows essentially no convergence. These plots will be discussed further in light of the research questions in the next chapter.

Standard Deviations and Standard Errors

The standard deviations of final round cut points for each condition group were calculated by taking the cut points for each individual panelist produced from individual logistic regressions and estimating the standard deviation. These results are presented in table 4.10 (following page). Condition 1, the control group, had the lowest standard deviations across cut points, Conditions 2 and 3 had similar standard deviations, and Condition 4 and 5 had the highest standard deviations.

Table 4.10

Cut Points and Standard Deviations of Cut Point for Final Round of Ratings

Achievement Level	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
Beginning to Developing	8.88 (0.69)	9.42 (1.22)	7.71 (1.25)	6.59 (2.07)	NA
Developing to Proficient	17.26 (0.73)	16.06 (1.48)	17.03 (1.58)	15.79 (2.24)	15.85 (2.63)
Proficient to Distinguished	22.49 (0.36)	24.76 (1.05)	24.98 (2.07)	NA	22.77 (2.64)

**Note.* Standard deviations are numbers inside parentheses.

Standard deviations and standard errors of the total number of attributes mastered of profiles categorized in each proficiency level for each round of ratings were calculated for each condition group. The complete set of these values is presented in Appendix M and the results for the final round of ratings are presented in Table 4.11. As with the convergence plots, there is greater variation for the Developing and Proficient cut points and less variation for the Beginning and Distinguished cut points for most of the conditions. For Condition 4 and Condition 5 that is not the case for the proficient and developing learner cut points. This pattern of standard error is likely due to the bounding of the ratings discussed above. The range of standard errors of final round ratings for all conditions is from 0.15 to 0.62 with an average of 0.31 and median value of 0.26. These values are lower than what was reported by Clark et al. (2017) and indicate that there was low overall variability in the final panelists' ratings.

Table 4.11

Standard Deviations and Standard Errors of Total Mastery Level for Final Round Ratings by Proficiency Level and Condition

Achievement Level	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
Beginning	1.20 (0.17)	1.32 (0.25)	1.44 (0.24)	1.48 (0.25)	NA
Developing	3.72 (0.46)	2.53 (0.34)	4.16 (0.47)	4.71 (0.62)	1.38 (0.25)
Proficient	2.76 (0.34)	2.55 (0.38)	3.61 (0.42)	1.47 (0.19)	3.07 (0.38)
Distinguished	0.97 (0.15)	0.99 (0.22)	1.37 (0.22)	NA	1.31 (0.26)

**Note.* Standard errors are numbers inside parentheses. Conditions 4 and condition 5 did not have ratings for Distinguished Learners or Beginning Learners respectively. The standard deviation ranged from 0.97 to 4.71 with a mean of 2.22 and median value of 1.48. The range of standard errors for all conditions is from 0.15 to 0.62 with a mean of 0.31 and median value of 0.26.

Nonparametric Longitudinal Factorial Analysis

A nonparametric longitudinal factorial analysis using the nparLD package in R was attempted as planned in the methods sections. First, data was prepared by estimating each of the cut points for each panelist for each condition and round of the standard settings. These cut points were used as the dependent variables of the analysis. Unfortunately, the omnibus analysis failed due to the unbalanced nature of the study design. While the method is capable of handling some missing data, the design of the study resulted in a situation where estimation was impossible with current methods. This stemmed from two main problems. First, the covariance matrix was singular which can produce instability in the computation of the regression coefficients and their standard errors, leading to inflated Type I error rates and reduced statistical power. This was due to a great deal of multicollinearity in the data because while some panelists changed their answers between rounds, many kept them the same. This meant that the data was essentially the same for some panelists in too many cells of the data. Furthermore, the difference in cut points between the range finding and pinpointing rounds (as visible in the plots in Appendix N) was substantial. Between these rounds, the logistic regression function was used to pinpoint a small range of choices for panelists. The available profiles went from profiles every three total attributes mastered to five consecutive levels of total attributes mastered for each cut point. This large shift in ratings in this round compared to the smaller shifts in ratings compared to the other rounds resulted in an unmanageable difference in the recommended cut points with regard to the multicollinearity of the data and fluctuations of the sphericity.

The second problem was due to the imbalance of the study design in terms of missingness. The cells for the lower and higher cut points of Condition 4 and Condition 5,

respectfully, were completely empty. These empty cells caused the estimation process to produce infinite values and for estimation to fail.

An exhaustive set of solutions were explored for these problems, including listwise deletion of data, transforming the data, imputing the missingness, and using a generalized linear mixed model approach, but in all cases, the extraordinary imbalance of the design of the study made finding a meaningful omnibus test infeasible. Instead, the data was subdivided into two groups to remedy the problem. One test was conducted using only data from Conditions 1, 2, and 3. These conditions did not have the empty cell challenges of Conditions 4 and 5. Then a second test was conducted with all conditions but using only the results from the middle cut point and ignoring the lower and upper cut points. This also remedied the problem because all condition groups include the developing to proficient learner cut point.

We conducted power analyses with the new designs to ensure there was sufficient ability to detect an effect. During the power analyses, we utilized the F test statistics for ANOVA with repeated measures with within and between factor interactions. Alpha was set to $\alpha = .05$ and beta was set to $\beta = .2$. Because no data is available to establish expected effect size, we used small, medium, and large effects of $f = .10$, $f = .25$, and $f = .40$, respectively, recommended by the G*power manual (Faul et al., 2009). For the test with only Conditions 1, 2, and 3, we found that the required sample size to detect a small sample size was 18. This means that adding one more panelist to each group would have been sufficient to detect an effect for this test. For medium effect sizes, we need nine participants and for large effect sizes only six total participants. This indicates that unless the effect is small, we should be able to detect it using this new test and the current sample size. For the second test using only the middle cut point, we again found that we would need 30 participants to obtain the targeted power. That said, for

medium or large effect sizes the analysis stated we would need only ten total participants for each. These results indicate that if no significant effect is found and the effect size is small, we may fail to find a significant result if there is. In this regard, feasibility limits our study, however, if the effect size is medium to large, we should have plenty of power to find a significant result. The nonparametric repeated measure factorial ANOVA using only results from condition 1 to 3 revealed a nonsignificant main effect of the condition group factor on the resulting cut point $F(1.566, 7.719) = 0.760, p = 0.438$. The degrees of freedom are adjusted using a Greenhouse-Geisser Adjustment to account for the heteroscedastic ranking of observations (Noguchi et al., 2012). When using an ANOVA-type statistic (ATS), it is recommended (Bathke et al., 2009) that the degrees of freedom for the whole plot factor (condition group) is $F_{(\hat{f}, \hat{f}_0)}$ but the degrees of freedom for the subplot factors (cut point and round of rating) are $F_{(\hat{f}, \infty)}$ because the ATS becomes too conservative for testing an effect with sub-plot factors on a finite denominator degrees of freedom. These results suggest that there is not a significant difference between Condition 1, 2, or 3. The results of the analysis are summarized in Table 4.8. Because the results of the factor of interest were not significant, follow-up contrasts were not conducted.

Table 4.8

Results of Nonparametric Factorial ANOVA for Conditions 1, 2, and 3

	ATS	Degrees of Freedom	<i>p</i> -value
Condition	0.760	1.566, 7.719	.469
Cut Point	258.650	1.62	<.001
Round	1.742	1.807	.1785

The nonparametric repeated measure ANOVA using only results from the middle cut point revealed a significant main effect of the condition group on the resulting cut point with $F(2.086, 8.950) = 6.895, p = 0.015$. This suggests that the alterations of elements of the standard setting across condition groups did have an effect on the resulting cut points. The main effect of the round, $F(1.300) = 0.459, p = 0.548$, and the interaction of the condition and the round, $F(2.606) = 0.413, p = 0.715$, were nonsignificant.

The relative marginal effects of the condition group over the rounds of rating are presented in Figure 4.2 on the follow page with confidence intervals. The large differences in the marginal effects of Condition 4 and Condition 5 compared to Condition 1 are easily seen in the vertical dislocation of the lines and confidences intervals.

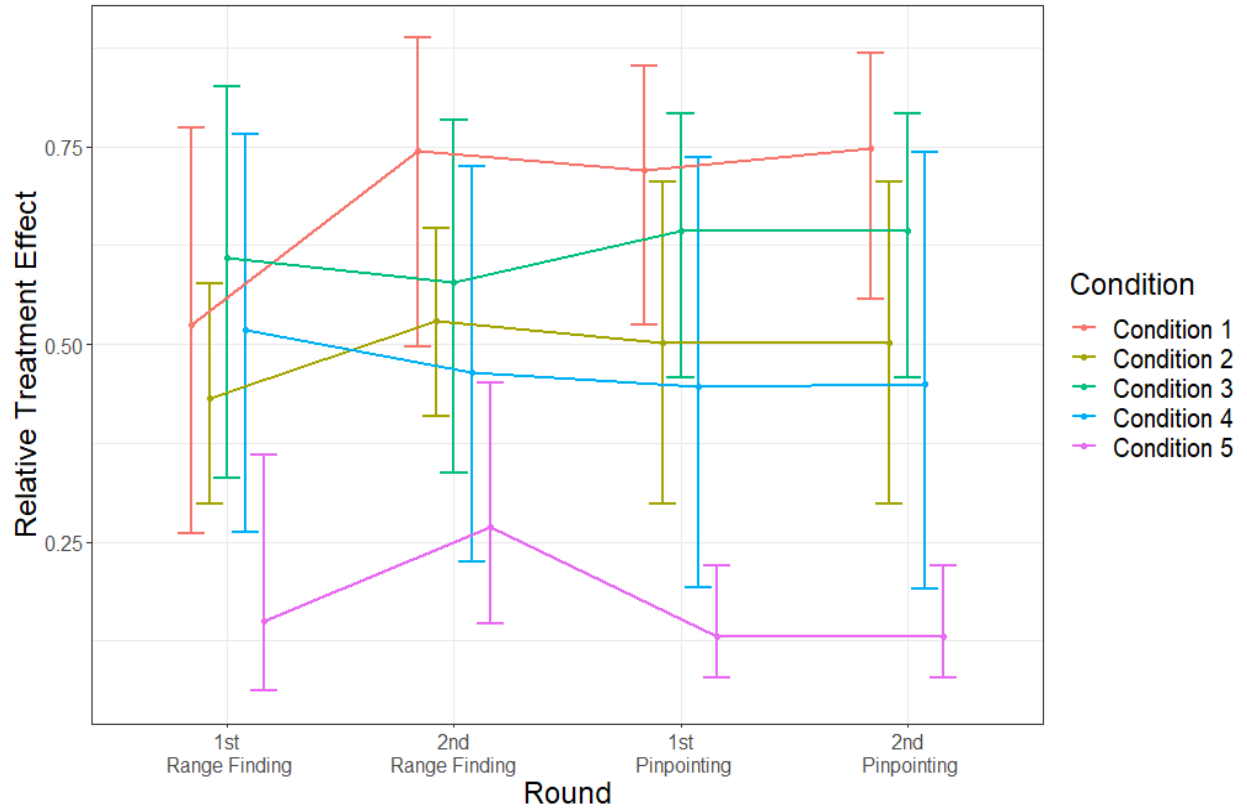
Table 4.9

Results of Nonparametric Repeated Measure ANOVA for the Middle Cut Point

	ATS	Degrees of Freedom	p-value
Condition	6.895	2.086, 8.950	0.015
Round	0.459	1.300	0.548
Interaction	0.413	2.606	0.715

Figure 4.2

The Relative Marginal Effects of Treatment Conditions for Middle Cut Point Ratings



Because of the significant main effect of the condition group on the resulting cut points, planned contrasts were conducted to compare the various groups as described in the methods section using the `nparscomp` package in R (Konietschke et al., 2015). The contrasts matrix used for the planned contrasts was

$$C = \begin{pmatrix} c'_1 \\ c'_2 \\ c'_3 \\ c'_4 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & .5 & .5 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

which indicates a comparison of the of Condition 1 to Condition 2, Condition 1 to Condition 3, Condition 1 to a combination of Condition 4 and Condition 5, and Condition 4 compared to

Condition 5. Using the multiple contrast test procedure (MCTP; Konietschke et al., 2015), a rank-based estimator \hat{p} of the unweighted treatment effects was estimated for each contrast with p -values that control the family wise error rate. The results of the contrasts are presented in Table 4.10. In line with the results of the first ANOVA described above, the results of the first and second contrast have p -values of 0.067 and 0.927, respectively, indicating that any differences between Condition 1 and Condition 2 and between Condition 1 and Condition 3 are not statistically significant. However, the third contrasts, with a treatment effect estimator $\hat{p}_3 = 0.164$, $t = -4.642$, and $p < .001$ indicated that there was a significant difference between the control group and the combined effects of Conditions 4 and 5. This suggests that the range of the profiles presented to panelists has an impact on the resulting cut points. In addition, the fourth contrasts, with a treatment effect estimator $\hat{p}_4 = 0.219$, $t = -2.733$, and $p = .002$ indicates that there was a significant difference between Condition 4 and Condition 5 groups. At 0.219, the effect of being in Condition 4, the reference in this comparison, as opposed to being in Condition 5 is considered moderate. This suggests that the range of profiles presented to panelists has an impact on the resulting cut points.

It can be noted that across the comparisons the effect size estimators have a unique trend in regard to the p -values in that they appear to be correlated, with larger estimates related to larger p -values. For the first contrast, the effect size for being in Condition 2 verses being in the reference group, Condition 1, is 0.258, and the p -value is just above the threshold of significance at 0.067. Then in the second comparison the estimator is larger with an effect of 0.438 and much larger p -value of 0.972. Furthermore, the smallest estimator is associated with the third comparison which also had the smallest p -value, and the fourth contrast had a moderate effect size and significant p -value of 0.002. This pattern generally contradicts expectations where we

would expect to see a larger effect size associated with a smaller p-value, but it is likely indicative of the dispersion of the data points and variations in sample sizes across the conditions and contrasts. As noted above, Condition 2 had the smallest variation and dispersion, but the ratings for the profiles that Condition 1 saw that overlapped with Condition 2 had high agreement (a Fleiss' kappa of .86). This manifests in a moderate effect size and a lack of significance. Then, for the second comparison, we noted that the lack of change in ratings and the greater variation within Condition 3 resulted in less commonality with Condition 1, however the resulting cut points were nearly identical. This resulted in there being a large but noisy effect that lacked significance. Next, when considering the third contrast, the sample size for that contrast was 14 (five from Condition 1 and Condition 4 and four from Condition 5), which made it easier for that contrast to create a significant value, while the effect size may have been lessened because the effects of Condition 4 membership (that skewed ratings negatively) and the effects of Condition 5 membership (that skewed ratings positively). Finally, the contrast between Conditions 4 and 5 had the most consistent dispersion and the greatest difference in final ratings, resulting in a moderate effect size (larger than that of the third contrasts, for example) and small significance.

Table 4.10

Results of Planned Contrasts of Panelist Middle Cut Points

Comparison	Estimator	Lower	Upper	Statistic	p-value
c'_1	0.258	0.103	0.512	-2.354	0.067
c'_2	0.438	0.230	0.669	-0.648	0.927
c'_3	0.164	0.076	0.318	-4.642	<.001
c'_4	0.219	0.082	0.469	-2.733	0.002

Panelist Exit Survey Ratings

The panelist exit survey given at the end of the standard setting provided mostly positive feedback and agreement with the results and process of the standard setting. During the standard setting exit survey, panelists were asked if they agree with the final cut point of the survey for each of the achievement levels, and if they did not agree to provide a recommendation for the final location. The results of the survey are presented in Appendix O. For the Developing to Proficient cut point, only one panelist thought the cut point, set at 17 by their condition group, was too low by three points. For the Beginning to Developing cut point, only one panelist thought the cut point, set at nine by their condition group, was too high by one point. And for the Proficient to Developing Learner cut point, one panelist believed the cut point, set by their group at 24, was too low by three points, and one panelist in a different condition group thought the cut point, set by their group at 24 was too high by two points. Despite these four panelists that wanted minor adjustments, most panelists agreed with the resulting cut points.

In addition to their agreement with the specific cut points, panelists were asked a series of questions about their belief about the standard setting process. Across panelists the results were overwhelmingly consistent and positive. Panelists stated that they believed the standard setting goals were met, considered the ratings process clear and fair, that they would defend the resulting cut points, and that participating in the standard setting was beneficial to their understanding of assessment. Furthermore, Appendix P presents a series of comments about the standard setting. These ratings and comments are discussed further in the following chapter.

Chapter 6. Discussion

This chapter contains a discussion of the study questions considering the results. It draws conclusions for each question and then some general conclusions, considers study shortcomings, and discusses next steps for research.

Discussion by Question

Here is a brief discussion of each question considering the results.

Question 1: How Does the Set of Profiles Presented to Panelists During a DCM Standard Setting Impact the Resulting Cut Point?

To answer question one, we altered Condition 2 to have fewer cut points than the control group. Rather than having three profiles at each total attribute mastery level, Condition 2 saw only two profiles at each total attribute mastery level. The results of the nonparametric, repeated measure ANOVAs comparing Condition 2 to the control group, using both the full data set and using only the middle cut points data, did not produce a statistically significant result. This means that reducing the number of profiles viewed by participants did not significantly impact the resulting cut points.

Further examining the ratings of Condition 2 shows some interesting patterns. The first stage of the range finding of Condition 2 shows a smaller range than that of the control condition. This is most evident in the interrater agreement of profiles. With a Fleiss' Kappa of 0.87, Condition 2 had higher interrater agreement than the control (0.55) and all the other conditions. This greater degree of agreement is also evident in the convergence plots of

Condition 2 where the bars and whiskers for the range finding stages of the developing and proficient cut points point are not balanced and the minimum and maximum values are much narrower than for the Condition 1. The distribution is so imbalanced that the third quartile of the categorizations for Condition 2 is also the highest rating, and the median values are also the first quartiles. This suggests that the ratings are less smoothed and more condensed over the potential range of total attributes mastered in this condition group compared to the control. This is also evident in the standard deviations of the ratings of Condition 2 compared to Condition 1, which tended to be smaller for each round and proficiency level (see Appendix M). It seems that the additional third profile rating that Condition 1 made, which was the third most prevalent profile at that mastery level, caused the distribution of ratings to be more varied and smoother over the range of total attributes mastered compared to Condition 2. Importantly however, through the process of the standard setting, the ultimate cut points produced by the conditions are comparable to each other, only differing by one point (after rounding) for the Proficient to Distinguished Learner level but being the same for the other two cut points. The facilitator did not note any significant difference between Condition 2 and the control group during the standard setting regarding the level of discussion or participation. It seems that reducing the number of profiles presented to panelists, and therefore reducing the cognitive load and time required from panelists, may produce slightly less smooth distributions of ratings in the initial rounds of the standard setting but will have negligible effects on the resulting cut points.

Question 2: How Does the Variability of the Profiles Presented to Panelist During a DCM Standard Setting Impact the Resulting Cut Point?

To answer the second question, the panelists in Condition 3 were provided with variety in the set of profiles they saw. Instead of all panelists in this condition seeing the same profiles at each total mastery level, each panelists in Condition 3 saw a different set of three profiles at each total mastery level. The selection process of these profiles is presented in Table 3.1 and discussed in detail in that section. The results of the nonparametric, repeated measure ANOVAs comparing Condition 3 to the control group, using both the full data set and using only the middle cut points data did not produce a statistically significant result. This means that increasing the variability of profiles viewed by panelists did not significantly impact the resulting cut points.

That said, there was more variability, less agreement, and less convergence for Conditions 3. The interrater agreement for Condition 3 was the lowest for three out of the four rounds of rating based on Fleiss' Kappa (the pinpoint round of Condition 5 was only 0.01 Kappa lower). Comparing the convergence plots of the control and Condition 3 shows that Condition 3 had less variation over the rounds of standard settings but more variation within the rounds. This was because during the discussions between the first and second range finding and first and second pinpointing, panelists rarely (for only two ratings) changed their ratings. That said, comparing the standard deviations of the ratings of the control and Condition 3 shows that Condition 3 had more variation for ratings of all but one cut point (Condition 1's first round range finding was 0.10 higher though with a lower standard error). This increased degree of variation of ratings seems likely to stem from the wider range of profiles presented to panelists.

The facilitator noted that the variation of the profiles led to a great deal more discussion. As with other conditions, after the first range finding and first pinpointing rounds the facilitator collected the ratings of the group and presented it to the panelists with measures of variability. The facilitator asked individual panelists to explain their ratings on those profiles that had the greatest variation and presented the profile to the group. This led to a great deal more discussion in Condition 3 than in other condition groups. Panelists were able to draw distinctions between the profiles that they and their peers saw, developed more precise weightings of specific standards, and more thoroughly compared the paradigms they were each using to make the categorizations.

While Condition 3 had greater discussion than the other conditions, panelists in Condition 3 rarely changed their responses to come to greater consensus. This is evident in the stagnant convergence plots and Condition 3 having the lowest change in Fleiss' Kappa between rounds. This may have been because panelists knew that their set of profiles were different than the other panelists in the condition. Despite the lack of changing ratings, the panelists in this round ultimately produced final cut points nearly identical to the control. The Beginning to Developing Learner cut point was one lower than the control (nine compared to eight), the Developing to Proficient Learner cut point was identical to the control (17), and the Proficient to Distinguished Learner cut point was one higher than the control (25 compared to 24). The use of variability in profiles was successful in spurring discussion but was detrimental to building consensus. This is evident in the fact that the standard errors of the ratings for this group were larger for every rating category than the control. Despite the lack of consensus building and greater variability of ratings, the condition ultimately produced final cut points similar to the control. It may be beneficial to use variation of profiles in earlier rounds of standard settings to

stimulate more conversation, and then remove that variation of profiles in later rounds to build consensus.

Question 3: How Does Presenting Only Part of the Total Attribute Mastery Range to Panelists Impact the Resulting Cut Point Location?

To determine the effect of presenting only part of the total attribute mastery range to panelists, Condition 4 and Condition 5 were given profiles from only the lower two-thirds or upper two-thirds of the mastery range, respectfully. The results of the nonparametric, repeated measure ANOVA using the middle cut points for all groups produce a statistically significant result. The follow up contrasts showed a significant difference between both the combined effects of being in Conditions 4 and 5 compared to the control (third contrast), and a significant difference between Conditions 4 and Condition 5 (fourth contrast). This suggests that changing the range of profiles viewed by participants does have a significant impact on the resulting cut points.

Examining the convergence plots of these conditions shows that for the first round of ratings, the range between the lowest and highest ratings were larger than for the control group. It seems that panelists may have set their original ratings more widely due to the change in the range. Furthermore, the variance of the ratings as indicated by the standard deviations were consistently higher for Conditions 5 and Condition 5 than for Condition 1 on the middle cut points. It seems clear that the panelists' ratings were bounded by the ratings on the upper and lower cut points they were able to make but shifted ratings for the middle cut point.

That said, the resulting cut points for these conditions were different for the lowest cut point but relatively similar for the middle and upper cut points. The Beginning to Proficient

Learner cut point for condition four set at seven was two below that of the cut point set by the control for that achievement level. At 16 total attributes mastered, the Developing to Proficient Learner cut point for both Condition 4 and Condition 5 were one below the Developing to Proficient Learner cut point of 17 set by the control. Finally, the Proficient to Distinguished cut point of Condition 5 was 23, set one below that of the control. This variation suggests that panelists' ratings for middle ranges are resilient to alterations of the scale presented to them because of the process of the standard setting, but that the outer cut points may be more widely effected. The facilitator noted that a large portion of this resilience may be due to anchoring of the panelists in the current assessment regime. Panelists often referenced the need for knowing certain content or standard for the current standardized tests. For example, they one panelist stated that "Geometry only makes up 10% of the [current assessment system], so it's not as important fractions." This suggests that they had developed clear anchoring of the requirements for proficiency and were using these frameworks to make their categorizations and maintain resilience to changes in the scale.

Question 4: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Panelists'; (a) Confidence in Their Ratings, (b) Agreement with the Resulting Cut Point, and (c) Perception of the Fairness and Validity of the Cut Point Setting Process?

The responses of the exit survey provide few conclusions about the effects of the alterations of the conditions of the standard setting on panelists confidence in their ratings, agreement with the cut points, and perception of the fairness and validity of the cut point setting process. This is because most of the responses were the same (see Appendix O). After four and a half hours of standard setting, panelists seemed ready to be done and three out of five groups

answered all “Strongly Agree” for every question of the survey. Condition 4 had two panelists that selected “Somewhat Agree” for every question. Condition 3 also had one panelist answer all “Somewhat Agree” as well as one panelist answer all “Strongly Disagree.” This last panelist was flustered by the process and had difficulty entering the ratings in the survey form.

The panelists also had an opportunity to leave a comment at the end of the standard setting. These comments, presented in Appendix P, are overwhelmingly positive. Panelists stated that they appreciated hearing the perspectives of other educators, enjoyed learning about the DCM assessment system, complimented the facilitators explanations and leadership of the study, and overall gained from involvement as a professional development experience.

Question 5: How Do the Conditions of the Standard Setting Described in Questions 1 to 3 Impact the Observed Participation of Panelists?

The observed participation of the panelists was recorded in the post standard setting facilitator notes and is discussed here. The major difference in level of participation across the groups of the standard setting was that the third group had considerably more discussion than the other groups. The facilitator noted that the variation of the profiles in this group resulted in the need for more discussion and pushed the panelists to more deeply examine the process by which they made categorizations. That said, the facilitator also noted that this did not seem to help bring the panelists to consensus.

Another difference between the groups was the time required to complete the categorizations and come to consensus. Condition 2 took noticeably less time to make the ratings of the profiles. This was probably because of the reduced set of profiles they viewed meant less time needed to rate and an easier time coming to consensus during the discussion

phases. Conversations tended to be shorter and more easily developed agreement. This makes sense when considering the interrater agreement of Condition 2.

One major challenge of the standard setting was ensuring that all panelists logged on at the right time and were fully engaged throughout the standard setting. This was due in large part to the fact that the standard setting was held virtually. Panelists were asked to have a quiet place without interruption, but some were interrupted a few times. Furthermore, there were times when data had to be prepared and manipulated by the facilitator and the panelists were engaged, such as before the discussions and for the two points when logistic regressions were conducted. During this time, panelists were told that a short break would ensue. Ensuring both that the analysis was complete and that the panelists returned on time was difficult and caused short delays. This did not impact a particular condition group excessively, but when considering the observed participation of panelist, it is worth noting that the setting and transitions of the standard setting played an important role.

Conclusion

The greatest conclusions that should be drawn from this study is that standard settings produce arbitrary results that are hard to defend and that it is time for the federal mandate of accountability testing that produces three levels of proficiency to end. The methods used in standard settings for both CDMs and IRT are meant to push educators towards consensus. We see in Conditions 1, 2, and 3 of this study that these methods are effective. Even with fewer profiles or entirely different sets of profiles, the methods resulted in similar cut points. It is more the procedure of the standard setting than the actual rating of the profiles that made a difference.

Psychometricians and policy makers have relied on this consensus building mechanism to argue for the defensibility of the cut points but the results of this study, as well as several parallel studies in IRT, indicate that the cut points are inherently arbitrary and relatively meaningless. In our study, the arbitrary nature of the cut points is evident in the significant effect that the range presented to panelists in Conditions 4 and 5 had on the resulting ratings and cut points. That the resulting ratings and cut points were affected by the range indicates that there is no objective definition of what proficient fourth grade math achievement looks like. The cut points produced by the standard setting methods are simply downstream procedures of the standards writing process arbitrarily decided by academics and policymakers. Similar effects were noted in related studies of IRT standard settings methods such as the Angoff method (Clauser, Harik, et al., 2009; Clauser, Mee, et al., 2009) and Bookmark method (Clauser et al., 2017).

The standard settings are relatively meaningless in the sense that there are other sets of standards that could have been chosen with the same degree of defensibility. As an example, imagine a fourth grade mathematics teacher that spends time teaching their students to "solve real world problems involving multiplication of fractions and mixed numbers" using word problems (Georgia Department of Education, 2016, p. 40). Multiplication of fractions is a fourth grade standard and the fifth grade standards extend the fourth grade standard to include multiplication of fractions with word problems and real world problems. The 4th grade teacher may justifiably be trying to make the fractions more tangible to their students or have some social or cultural reason to logically choose to use word problems and real-world examples in their instruction. However, in doing so the teacher would be teaching a standard that is not on grade level. This instruction would utilize time that could be spent learning the other arbitrarily chosen standards that are "on grade level" that would better prepare students for the test. The

students could have mastered the standards in the next grade level in a certain domain but could be deemed not proficient on the test because they do not cover a sufficient number of standards across all domains in the their current year. The resulting proficiency designation in this case is relatively meaningless in comparison to what the students may actually know.

This situation highlights what is taken for granted by psychometricians; that the tests both limit what can be taught and yet the tests cannot cover all the material of the course. In most cases, the course content is condensed to a form that fits a multiple-choice item of the test even though not all KSAs translate equally well to multiple-choice items. In all cases, the course content is reduced to a set of items that only partially cover the content or the skills of the course. Even if a test has a high relative degree of content coverage, it is inherently limited. The tests in this sense are a virtual manifestation of Bentham's panopticon. Bentham, an 18th century philosopher and social theorist, devised a system in which a single security guard could monitor an entire prison by placing the guard in the middle of the prison with a one-way mirror between the guard and the prisons. The prisoners are coerced to behave by the fear that the guard could be watching them at any time. So too teachers and students are coerced into covering the state mandated standards to perform well on the test, not knowing specifically what questions and content will be asked.

Multiple standard setting experts were consulted in preparation for this study. In the words of one of the leading psychometricians in the field responsible for overseeing dozens of standard settings across multiple states, "You don't want to see how the sausage is made." This message was the same from others as well and was meant to express the messiness and lack of objectivity to setting cut points. In the five standard settings conducted for this study, every effort was made to remain an objective facilitator. But as the professionals in the field expressed

and as was clear during the standard settings, it would be easy to guide the panelists to reach a certain cut point, say for a state that wanted to lower the threshold of proficiency to improve results. The ambiguity of the process is compounded by the fact that many states have not conducted standard settings in years, equating and linking entirely new tests, some based on new standards, year after year. The whole process is like sausage making in that whatever we put in is unclear, it is all a little messy, and it is easy to form what comes out. Psychometricians collectively know that it is arbitrary and have a moral responsibility to speak up and work to change the laws because the results of accountability tests have real world implications for students, teachers, schools, and communities.

The simple solution is that the federal government remove the mandate that test create proficiency designation and instead regulate only the quality of assessments, the presentation of assessment data, and the use of assessment results. Like with car safety in which the federal government regulates the safety and quality of the car without mandating an exact model, educational assessment regulation should require that quality measurement tools be made available to students and educators without requiring those measures to produce specific designations of students. Arguments could be made for the use of graduation tests, entrance exams for specific programs or schools, or qualification of special services, but the yearly designation of students in third through eighth grade into proficiency levels serves no meaningful educational purpose and should be discontinued.

This study was initiated to find methods of bridging the gap between serving the purposes of accountability testing and pedagogically relevant testing. Political conditions make it doubtful that the federal accountability requirements will change any time soon. In the interim, CDMs

hold the promises for relieving the distraction that accountability testing creates while providing meaningful, pedagogically relevant assessment tools to students and educators.

Study Limitations

Conducting five standard settings produced a depth of knowledge about the standard setting process and ways to improve that process that are provided here as a discussion of the study shortcomings. To begin, the process of the standard settings was meant to build consensus among the panelists between the successive rounds of ratings. During data collection, a few steps could have been taken to provide for greater consensus building. First, the amount of time allotted to discussion was likely insufficient. We believed that panelists would not be willing to sit for more than four hours without increasing the honorarium (something that was not feasible) and so we limited the discussion portions to 15 and 20 minutes. During this time, the facilitator attempted to pick profiles that provided the greatest disagreement among panelists ratings and ask panelists how they made their categorizations. The limited amount of time allowed for discussing only three or four such profiles. Panelist sometimes stated that they were convinced during the discussion to change their responses and did so in the second and fourth rounds of rating. More time given to discussion would likely have resulted in more rating changes and greater consensus building among panelists between rounds. Furthermore, panelists were given the chance to update their profiles after the discussion portion. There is no reason they could not have been allowed to change their ratings during the discussion sections, but they were not instructed to do so. It is likely that some of the panelists in real time would have changed their ratings, but ultimately did not because they had to wait to make the changes until the designated time and so forgot or otherwise failed to make the changes.

There was likely an important impact of holding the standard setting in a virtual setting rather than an in-person setting. While every effort was made to ensure the materials and processes were the same for each panelist, variations in their environment likely impacted the resulting ratings. While few, some panelists were interrupted by events at their individual locations such as children and pets. Furthermore, the level of communication might have been greater in person. Video conferencing is a great tool but is less personal and natural than in-person conversation, particularly when meeting for the first time as the panelist were.

A final point of improvement and recommendation would be to improve the fluidity of the data analysis portion of the standard setting. During the data collection before the discussion and during the logistic regression portions after the second range finding and pinpointing rounds, data had to be copied and formatted from the online webform to the excel for presentation and R for analysis. While this process was practiced several times before the standard settings, it likely could have been improved further. The down time for these sections was only a few minutes, but some panelist disengaged. This likely would have been remedied either by being in person or by improving the methods.

Future Research

This study examined a limited set of characteristics about the standard setting process for diagnostic profiles, but there are many places for future research. One place for future research is developing a method of standard setting for diagnostic profiles that does not use the total number of attributes mastered as the underlying scale. Instead, an alternative decision process could be introduced that maintained the influence of the individual attributes. This could be done by weighing the attributes, developing a decision tree, or some other method.

Another place for future research is analyzing the impact of adding information about untested attributes to the profile ratings. In our study, the students were labeled as master or nonmaster of the attributes. However, because the attributes are measured by individual assessments, students could be labeled as nonmaster even though they had not yet taken the assessment. In cases where a student is untested, there is a degree of certainty around their true mastery status. How panelists would categorize profiles when students are untested is important. Consider for example the fact that classroom teachers assign the DCM assessments, so a student at the end of a year could be designated nonmaster of attributes by default without the assessment system actually having information on their mastery status.

A final point of research that is mentioned here that is worth examining is the impact of grainsize on the standard setting. Grainsize is the level of specificity and detail at which attributes are measured. Some attributes could be broken down into smaller attributes with a more refined grainsize or combined with other attributes to make larger grainsize. For example, in our study, standard 21 reads, “Know relative sizes of measurement units within one system of units including km, m, cm; kg, g; lb, oz.; l, ml; hr, min, sec.” (Georgia Department of Education, 2016, p. 33). This attribute could be broken into three attributes of measurement unit for weight, time, and size. While this study manipulated the range of the attributes presented to panelists, further research could manipulate the grainsize.

References

- Angoff, W. (1984). *Scales, Norms, and Equivalent Scores*. Educational Testing Service.
- Baldwin, P. (2021). A Problem with the Bookmark Procedure's Correction for Guessing. *Educational Measurement: Issues and Practice*, 40(2), 7–15.
<https://doi.org/10.1111/emip.12400>
- Bathke, A. C., Schabenberger, O., Tobias, R. D., & Madden, L. V. (2009). Greenhouse–Geisser Adjustment and the ANOVA-Type Statistic: Cousins or Twins? *The American Statistician*, 63(3), 239–246. <https://doi.org/10.1198/tast.2009.08187>
- Bradshaw, L. (2017). *A Classroom-embedded Innovative Assessment System for Learning and Accountability*.
- Brunner, E., Dette, H., & Munk, A. (1997). Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association*, 92(440), 1494–1502.
<https://doi.org/10.1080/01621459.1997.10473671>
- Brunner, E., & Puri, M. L. (2001). *Nonparametric Methods in Factorial Designs*. 53.
- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, 39(3), 253–263.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27(2), 145–163. <https://doi.org/10.1111/j.1745-3984.1990.tb00739.x>

- Chang, L. (1999). Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting Methods. *Applied Measurement in Education*, 12(2), 151–165.
https://doi.org/10.1207/s15324818ame1202_3
- Clark, A. K. (2015). *2015 Integrated Model Standard Setting: English Language Arts and Mathematics*. 198.
- Clark, A. K., Nash, B., Karvonen, M., & Kingston, N. (2017). Condensed Mastery Profile Method for Setting Standards for Diagnostic Assessment Systems. *Educational Measurement: Issues and Practice*, 36(4), 5–15. <https://doi.org/10.1111/emip.12162>
- Clauser, B. E., Baldwin, P., Margolis, M. J., Mee, J., & Winward, M. (2017). An Experimental Study of the Internal Consistency of Judgments Made in Bookmark Standard Setting: Judgments in Bookmark Standard Setting. *Journal of Educational Measurement*, 54(4), 481–497. <https://doi.org/10.1111/jedm.12157>
- Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An Empirical Examination of the Impact of Group Discussion and Examinee Performance Information on Judgments Made in the Angoff Standard-Setting Procedure. *Applied Measurement in Education*, 22(1), 1–21.
<https://doi.org/10.1080/08957340802558318>
- Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*, 46(4), 390–407. <https://doi.org/10.1111/j.1745-3984.2009.00089.x>

- Etikan, I., Musa, S., & Alkassim, R. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1. <https://doi.org/10.11648/j.ajtas.20160501.11>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Georgia Department of Education. (2016). *Georgia Standards of Excellence: Mathematics*. Georgia Department of Education. <https://www.georgiastandards.org/Georgia-Standards/Pages/Math-K-5.aspx>
- Georgia Department of Education. (2018). *Georgia's Application for the Innovative Assessment Demonstration Authority: Under Section 1204 of the Elementary and Secondary Education Act (ESEA)*. Georgia Department of Education.
- Goodwin, L. D. (1999). Relations Between Observed Item Difficulty Levels and Angoff Minimum Passing Levels for a Group of Borderline Examinees. *Applied Measurement in Education*, 12(1), 13–28. https://doi.org/10.1207/s15324818ame1201_2
- Hambleton, R. K., & Plake, B. S. (1995). Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education*, 8(1), 41–55. https://doi.org/10.1207/s15324818ame0801_4
- Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage.

- Hartz. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities_Blending Theory with Practicality.pdf*.
- Huynh, H. (2000). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard settings*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Karantonis, A., & Sireci, S. G. (2006). The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement: Issues and Practice*, 25(1), 4–12.
<https://doi.org/10.1111/j.1745-3992.2006.00047.x>
- Konietschke, F., Bathke, A. C., Hothorn, L. A., & Brunner, E. (2010). Testing and estimation of purely nonparametric effects in repeated measures designs. *Computational Statistics & Data Analysis*, 54(8), 1895–1905. <https://doi.org/10.1016/j.csda.2010.02.019>
- Konietschke, F., Placzek, M., Schaarschmidt, F., & Hothorn, L. A. (2015). An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals. *Journal of Statistical Software*, 64(9), 1–17.
- Koretz, D. (2017). *The Testing Charade: Pretending to Make Schools Better*. University of Chicago Press.
- Leighton, J., & Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press.
- Lewis, Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A Bookmark approach*. In D. R. Green (Chair), IRT-based standardsetting procedures utilizing behavioral anchoring. Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

- Lewis, Mitzel, H. C., Mercado, R. L., & Schultz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). Routledge.
- Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, 13, 33.
<https://doi.org/10.14507/epaa.v13n33.2005>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 276–282.
<https://doi.org/10.11613/BM.2012.031>
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Lawrence Erlbaum Associates.
- Noguchi, K., Gel, Y. R., Brunner, E., & Konietzschke, F. (2012). Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *Journal of Statistical Software*, 50(12). <https://doi.org/10.18637/jss.v050.i12>
- Pellegrino, J. W., Jones, L. R., Mitchell, K. J., & National Research Council (U.S.) (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. National Academy Press.
- Perie, M. (2008). A Guide to Understanding and Developing Performance-Level Descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15–29.
<https://doi.org/10.1111/j.1745-3992.2008.00135.x>
- Resnick, D. P. (1980). Chapter 1: Minimum Competency Testing Historically Considered. *Review of Research in Education*, 8(1), 3–29.
<https://doi.org/10.3102/0091732X008001003>

- Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff Methods. *Alberta Journal of Educational Research*, 52(1).
- Shah, D. A., & Madden, L. V. (2004). Nonparametric Analysis of Ordinal Data in Designed Factorial Experiments. *Phytopathology*®, 94(1), 33–43.
<https://doi.org/10.1094/PHYTO.2004.94.1.33>
- Shepard, L. (2008). A Brief History of Accountability Testing, 1965-2007. In *The Future of Test-Based Educational Accountability* (pp. 25–47).
- Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting Performance Standards for Student Achievement*. National Academy of Education.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34.
<https://doi.org/10.1111/emip.12189>
- Skaggs, G., Hein, S. F., & Wilkins, J. L. M. (2016). Diagnostic Profiles: A Standard Setting Method for Use With a Cognitive Diagnostic Model. *Journal of Educational Measurement*, 53(4), 448–458. <https://doi.org/10.1111/jedm.12125>
- Skaggs, G., & Tessema, A. (2001). *Item Disordinality with the Bookmark Standard Setting Procedure*. 22.
- Wilson, M. (2018). Making Measurement Important for Education: The Crucial Role of Classroom Assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20.
<https://doi.org/10.1111/emip.12188>

Appendix A

Panelists Recruitment Letter

Dear Colleagues,

We invite your teachers or math specialists who have experience teaching 4th grade mathematics to participate in a research study. The study will include participating as a panelist in a study about how standards competencies relate to state assessment achievement levels. An honorarium of \$125 will be provided to participants completing the study.

We are in the process of developing a new standard setting approach for standards-focused assessment systems that can provide diagnostic feedback to teachers and students throughout the year while also being able to classify students according to achievement levels at the end of the year.

In this process, we need the help of educators with expertise in fourth grade mathematics to participate as panelists in what is known as a *standard setting*. In this standard setting, educators will relate profiles of standards that have been learned to different achievement levels based on their experience and expertise with standards and with students in their classroom. This study will conduct a standard setting utilizing the Georgia Standards of Excellence and the Georgia Achievement Levels: Beginning Learners, Developing Learners, Proficient Learners, and Distinguished Learners. This study is for research purposes only to understand how a process like this would work and is not being used by the state of Georgia.

We are asking for your help to either act as a member of the panel or to distribute this letter and the associated interest survey to experts you may know.

If you are a 4th grade math teacher or school mathematics specialist with familiarity or experience working with 4th grade mathematics students, please consider completing the associated interest survey. Alternatively, if you know of anyone that would meet these descriptions, please consider sharing the interest survey. We would be particularly grateful if you could distribute it to relevant listservs, discussion board, or similar platforms. Here is a link to the survey:

[Insert link to interest survey]

Participants will:

- Complete an independent, asynchronous training and survey (estimated time <30 minutes)
- Act as a panelists in a standard setting (estimated 4 hours)
- Receive \$125 in compensation for their time and effort
- Help to develop better assessment systems for the students of Georgia

The standard setting sessions are currently scheduled for:

- [insert times of standard setting]
- [insert times of standard setting]
- [insert times of standard setting]
- [insert times of standard setting]
- [insert times of standard setting]

Thank you for your assistance. If you have any questions, please do not hesitate to reach out to me, Zack Feldberg at feldberg@uga.edu.

Thank you,
Zack Feldberg, M.A.T.
PhD Candidate
University of Georgia

Appendix B

Panelist Recruitment Survey

Panelists Recruitment Survey

Start of Block: Contact Info

Q0 The following survey collects your contact info, demographics, experiences, and interests in acting as a panelists for a standard setting. Thank you for your responses!

Q1 First Name

Q2 Last Name

Q3 Email

Q4 Preferred Phone Number

End of Block: Contact Info

Start of Block: Demographics

Q5 What is your race?

- ☐ African American (1)
- ☐ American Indian/Alaska Native (2)
- ☐ Asian (3)
- ☐ Hispanic/:atino (4)
- ☐ Native Hawaiian/Pacific Islander (5)
- ☐ White (6)
- ☐ Prefer not to answer (7)

Q6 What is your gender?

- ☐ Female (1)
- ☐ Male (2)
- ☐ Prefer not to answer (3)

End of Block: Demographics

Start of Block: Experiences

Q7 What is your current role?

- ☐ Classroom Teacher (1)
 - ☐ Building Administrator (2)
 - ☐ District Staff (3)
 - ☐ State Education Agency Staff (4)
 - ☐ University Faculty/Staff (5)
 - ☐ Community Member (6)
 - ☐ Other (7) _____
-

Appendix C

Post In-Advance Training Survey

Start of Block: Block 1

Q0 The following surveys asks questions related to the in-advance training for the standard-setting. The questions will ask for understanding of the training information and provides you with space to ask questions you may have. Thank you for your participation!

End of Block: Block 1

Q1 Please rate your understanding or comfort with each of the following from Poor/Fair/Good/Excellent.	Poor (1)	Fair (2)	Good (3)	Excellent (4)
Characteristics of students who take the Georgia Milestones (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How prepared you feel to take part in the standard setting process (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The purpose of standard setting (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The 4th grade Mathematics Georgia Excellence Standards (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The 4th grade Mathematics Achievement Level Descriptors (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How to interpret a student profile (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The expectations for maintaining security of information (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How testlets measure the intended content (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How testlets are made accessible to students and teachers (9)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
What a student is expected to do during a Navvy assessment (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How Navvy results are reported (11)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Discussing achievement level descriptors with other panelists (12)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Discussing profiles with others (15)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q2 What remaining questions do you have about the standard setting process? Please do not hesitate to ask anything you may be unsure of. We will make sure to answer these questions in

our live session together.

Appendix D

Standard Setting Exit Survey

Standard Setting Exit Survey

Start of Block: Default Question Block

Q1 The following survey asks about your thoughts and opinions related to the standard setting. Your responses will not impact your participation or receipt of compensation for participation in the standard setting. Thank you for your participation in the standard setting and this survey.

Page Break

Q2 Considering the cut point (the total number of mastered attributes) that the group made between the "Developing Learner" and "Proficient Learner", which of the following do you agree with:

- ☐ It was too low (1)
- ☐ It was appropriate (2)
- ☐ It was too high (3)

Display This Question:

If Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too low

Or Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too high



Q2A

In the previous question, you chose the option $\{Q2/ChoiceGroup/SelectedChoices\}$ in reference to the cut point between the "Developing Learner" and "Proficient Learner."

Please enter the value you believe would be more appropriate and

Q3 Considering the cut point (the total number of mastered attributes) that the group made between the "Beginning Learner" and "Developing Learner", which of the following do you agree with:

- ☐ It was too low (1)
- ☐ It was appropriate (2)
- ☐ It was too high (3)

Display This Question:

*If Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too low
Or Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too high*



Q3A

In the previous question, you chose the option $\{Q3/ChoiceGroup/SelectedChoices\}$ in reference to the cut point between the "Begging Learner" and "Developing Learner."

Please enter the value you believe would be more appropriate and

Q4 Considering the cut point (the total number of mastered attributes) that the group made between the "Developing Learner" and "Proficient Learner", which of the following do you agree with:

- ☐ It was too low (1)
- ☐ It was appropriate (2)
- ☐ It was too high (3)

Display This Question:

*If Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too low
Or Considering the cut point (the total number of mastered attributes) that the group made between t... = It was too high*



Q4A

In the previous question, you selected $\{Q4/ChoiceGroup/SelectedChoices\}\{Q4/ChoiceGroup/SelectedChoices\}$.

Please enter the value you believe would be more appropriate and

Page Break

Q5 Please select the level of agreement you have with each of the following statements.

	Strongly Disagree (1)	Somewhat Disagree (2)	Somewhat Agree (3)	Strongly Agree (6)	Not Applicable (7)
The goals of the standard setting were met (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The In-advance training prepared me for the standard setting (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The training at the start of the standard setting prepared me for the standard setting (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The standard setting session was well organized (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was clear what knowledge, skills, or abilities were associated with a certain profile (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I considered the achievement level descriptors when rating each profile (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I considered other panelists opinions when I rated each profile (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I considered my experience when I rated each profile (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have enough time to complete the required tasks (9)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I understood how to complete the rating task (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel confident in the ratings I made (11)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I feel confident in the ratings the panel made (12)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The procedure for recommending cut points was free of bias (13)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If asked, I would defend the groups cut point for the distinction between "Developing Learners" and "Proficient Learners." (14)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If asked, I would defend the groups cut point for the distinction between "Beginning Learners" and "Developing Learners." (15)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

If asked, I would defend the groups cut point for the distinction between "Proficient Learners" and "Distinguished Learners." (16)

☐ ☐ ☐ ☐ ☐

Page Break

Q6 Click to write the question text

	Strongly Disagree (1)	Somewhat Disagree (2)	Somewhat Agree (3)	Strongly Agree (4)	Not Applicable (5)
Participating in the standard setting improved my understanding of assessment (1)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participating in the standard setting improved my understanding of the student population (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participating in the standard setting improved my understanding of diagnostic measurement (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participating in the standard setting improved my understanding of the diagnostic measurement based assessment systems (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, I valued my participation in the standard setting (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Page Break

Q7 Thank you for your participation in the standard setting. We'd like to invite you to leave any final comments you may have in the box below:

Appendix E

Standard Setting Agenda

In-Advance Training	
Estimated Less than 30 Minutes: Asynchronous	In-advance training and survey
Stage 1: In-Person Training	
9:00-9:30 AM	Training <ul style="list-style-type: none"> • Welcome, participant/facilitator introductions • Purpose • Processes • Achievement Level Descriptors • Questions from in-advance training survey • Practice standard setting • Time for final questions before standard setting
Stage 2: Range Finding	
9:30-10:15 AM	Independent range finding
10:15-10:30 AM	Collection of data and discussion
10:30-11:00 PM	Second independent ratings
Break	
11:00-11:30 PM	Lunch break Facilitator prepares materials for Stage 3
Stage 3: Pinpointing	
11:30-11:40	Provide panelists with list to update profile packet
11:40-12:10 PM	Independent pinpointing
12:10-12:30 PM	Collection of data and discussion
12:30-12:45 PM	Final independent ratings
12:45-12:50 PM	Make final cuts and present to panelists
12:45-1:00 PM	Exit survey

Note. Total expected time is four hours.

Appendix F

Standard Setting Facilitator Guide

Standard Setting Facilitator Guide

1. Two Weeks in Advance
 - 1) Mail all participants their profile packets.
 - 2) Contact each participant to ensure receipt.
2. Day before standard setting
 - 1) Email all participants a reminder email that instructs them to have an updated version of Zoom, an updated version of adobe or other PDF viewer, and the following resources:
 - i. The training material
 - ii. The ALD
 - iii. The example student profile
 - iv. The link for the zoom video
3. Preparation
 - 1) Ensure all materials are ready to include:
 - i. The in-person training PowerPoint
 1. The file with Policy ALDs
 - ii. Practice standard setting material
 1. The sample student profiles
 2. The sample range finding
 - iii. The student profiles for the specific condition
 - iv. Links to the Range finding Qualtrics survey
 - v. The survey for the Pinpointing (which will require adjustment)
 - vi. RStudio and the R code for the logistic regression
 - vii. Contact information for all participants and potential backup participants (incase a phone call or other contact is required)
 - 2) Get on Zoom 8:30 AM
4. Introductions and Training 9:00-9:30 AM
 - 1) Welcome participants, introduce yourself, and give a chance for everyone to introduce themselves.
 - 2) Provide panelists brief review of training by covering slides of on-site training to include:
 - i. Purpose
 - ii. Processes
 - iii. Roles and responsibilities
 - iv. Performance ALDs
 - v. Questions from in-advance training survey
 - 3) Lead discussion on ALDs
 - 4) Practice standard setting

- 5) Conduct a practice standard setting by instructing panelists to use the practice categorizing student profiles and placing their responses in the practice Qualtrics survey. Ensure that all panelists have:
 - i. The sample student profiles in the profile packets
 - ii. Link to the practice standard setting survey
- 6) Ask participants to share how they made the categorizations they did
- 7) Ask for any final questions before beginning the range finding
5. Range Finding
 - 1) 9:30-10:15 AM: Instruct panelists to independently conduct range finding. Tell them that they will get to a prompt that asks them to pause before submitting responses for discussion.
 - 2) 10:15-10:30 AM: Collect data into excel and present which profiles have the greatest disagreement. Facilitate discussion on why panelists rated profiles differently
 - 3) 10:30-11:00 AM: Give panelists and opportunity to conduct second ratings. Instruct them to submit responses via Qualtrics when done.
 - 4) Instruct panelists that they are on a lunch break and to return by 11:30 AM.
6. Data Analysis I
 - 1) Download Data
 - i. Go to relevant Qualtrics survey
 - ii. Click **Results**
 - iii. Click **CSV (Comma Separated)**
 - iv. Click **Export Pages**
 - v. Click **Download**
 - 2) Select Relevant Data
 - i. Under Rating Question, Copy and paste just the counts section
 - ii. Delete the percentage columns, ensuring that the labels above the count data stays in place
 - iii. Save file as **.csv** to Standard Setting Folder that also has R code.
 - 3) Analyze data
 - i. Open **Standard Setting R Code.R** in Rstudio
 - ii. Set working Directory to file location
 - iii. Run R code
 - 4) Create list of new profiles
 - i.
 - ii. Based on the logistic regression function produced from the rating of the range finding, find the level of total attributes mastered that provides a probability of .5 of being in any two adjacent proficiency levels. Create a list of profiles that are two below, the level of, and two above the total attributes mastered for this value.
 - iii. Distribute the list to panelists and ask them to remove all profiles not included in the list.
 - 5) Create new Qualtrics Survey

- i. Based on the p values, create a new Qualtrics study that contains only the profiles created in the list above.
 - ii. Distribute to panelists via email or zoom chat.
- 6) Return to zoom video by 11:25
- 7. Pinpointing
 - 1) 11:30-12:00 PM: Instruct panelists to review the new set of profiles and the new link to the updated Qualtrics survey for the pinpointing
 - 2) Instruct panelists to independently conduct the pinpointing. Tell them that they will get to a prompt that asks them to pause before submitting responses for discussion.
 - 3) 12:00-12:15 PM: Collect data into excel and present which profiles have the greatest disagreement. Facilitate discussion on why panelists rated profiles differently
 - 4) 12:15-12:45 PM: Give panelists and opportunity to conduct second pinpointing. Instruct them to submit responses via Qualtrics when done.
 - 5) Tell panelist there will be a 5-10 minute break before a final exit survey.
- 8. Data Analysis II
 - 1) Download Data
 - i. Go to relevant Qualtrics survey
 - ii. Click **Results**
 - iii. Click **CSV (Comma Separated)**
 - iv. Click **Export Pages**
 - v. Click **Download**
 - 2) Select Relevant Data
 - i. Under Rating Question, Copy and paste just the counts section
 - ii. Delete the percentage columns, ensuring that the labels above the count data stays in place
 - iii. Save file as **.csv** to Standard Setting Folder that also has R code.
 - 3) Analyze data
 - i. Open **Standard Setting R Code.R** in Rstudio
 - ii. Set working Directory to file location
 - iii. Run R code
 - 4) Present final cut points
 - i. Present the final cut points to the panelists.
- 9. Exit survey 12:50-1:00 PM
 - 1) Instruct the panelists to complete the exit survey.
 - 2) Thank the panelists for their participation.
- 10. Complete the Post Standard Setting Facilitator Notes Questions

Appendix G

Post Standard Setting Facilitator Notes

1. How would you describe the level of participation of the panelists?
2. How easy or hard was it to facilitate discussion among the panelists?
3. Were there any parts of the standard setting that were particularly memorable?
4. If any, what specific challenges did you run into?
5. If it has, how has discussion in this group differed from other groups?
6. If it has, how has the time required to complete the profile categorizations differed from other groups?
7. Based on the interactions and discussion, did the panel come to consensus easily or did they have more difficulty? Please provide specifics.

Appendix H

In-Advance Training Material

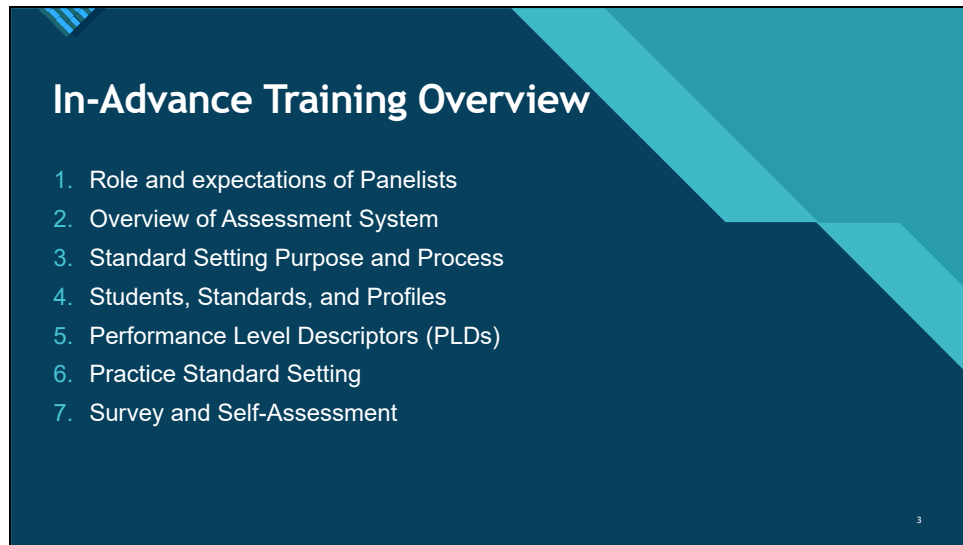
Slide 1



Slide 2



Slide 3



In-Advance Training Overview

1. Role and expectations of Panelists
2. Overview of Assessment System
3. Standard Setting Purpose and Process
4. Students, Standards, and Profiles
5. Performance Level Descriptors (PLDs)
6. Practice Standard Setting
7. Survey and Self-Assessment

3

This slide features a dark blue background with a light blue geometric design in the top right corner. The title 'In-Advance Training Overview' is in white. A numbered list of seven items is presented in light blue. A small number '3' is in the bottom right corner.

Slide 4



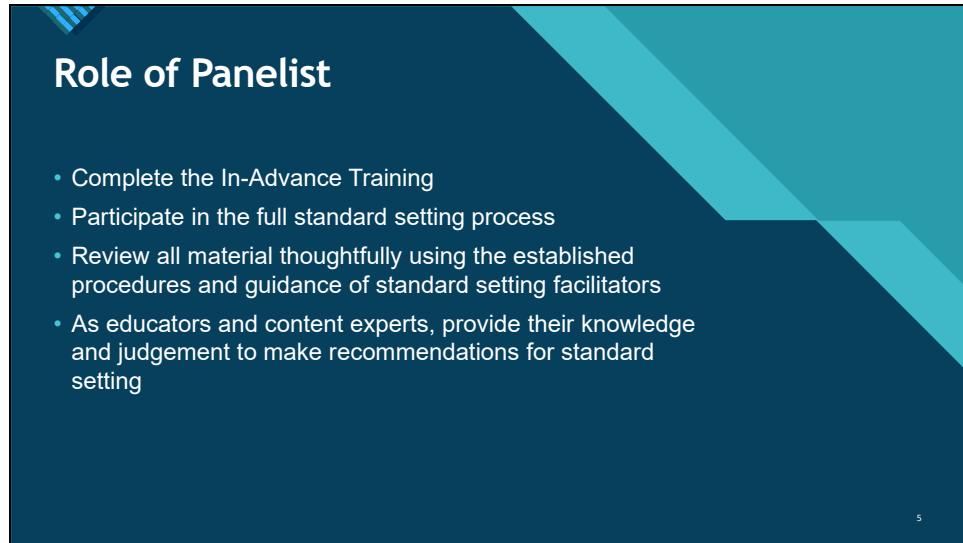
**In-Advance Training for
Standard Setting**

Roles and Expectations

4

This slide features a dark blue background with a light blue geometric design in the top right corner. The title 'In-Advance Training for Standard Setting' is in white. Below it, the subtitle 'Roles and Expectations' is in a smaller white font. A small number '4' is in the bottom right corner.

Slide 5


The slide has a dark blue background with a light blue geometric design on the right side. The title "Role of Panelist" is in white. The list of bullet points is in light blue.

Role of Panelist

- Complete the In-Advance Training
- Participate in the full standard setting process
- Review all material thoughtfully using the established procedures and guidance of standard setting facilitators
- As educators and content experts, provide their knowledge and judgement to make recommendations for standard setting

5

Slide 6

The slide has a dark blue background with a light blue geometric design on the right side. The title "Role of Panelist" is in white. The list of bullet points is in light blue.

Role of Panelist

- Act as a team member in courteous discussion when necessary
- Preserve the security of all material
 - Refrain from talking about specifics with anyone outside the setting
 - Do not reproduce or share any materials from standard setting
 - Maintain security of electronic devices
- Ask questions and provide comments or concerns to facilitators (Feldberg@uga.edu)

6

Slide 7

Expectations: Communicating and Confidentiality

Feel free to discuss:

- The Navy System of assessments
- Anything you learned about the process of meaningful educational assessment
- The academic content or skills related to the standards

Please do NOT discuss:


- Specific questions, items, testlets, test results, or content from the assessments

2

Slide 8

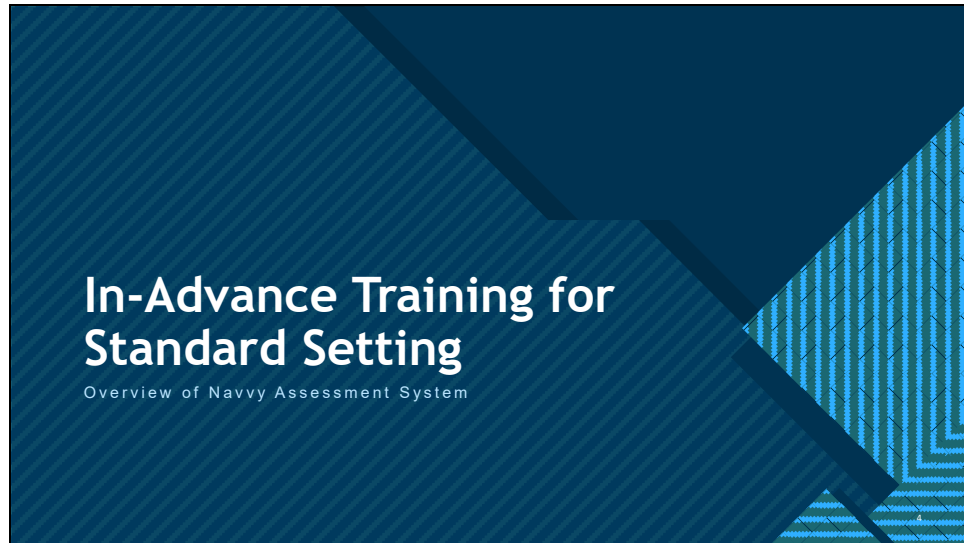
Expectations: Complete In-Advanced Training

- Module with video and several activities
- Complete at own pace
- Complete the self-assessment
- Ask any questions in the self-assessment
- Please complete the in-advance training before the panel setting meeting



3

Slide 9



Slide 10

The slide has a dark blue background with a light blue geometric shape on the right. The word 'NAVVY' is written in large, colorful, block letters (N: blue, A: yellow, V: red, V: green, Y: purple). To the right of the letters is a red line-art icon of a lightbulb with a hand inside it. Below the logo is a bulleted list of features.

- **Navvy (A Navigator)** helps navigate what students know, do not know, and need to learn.
- Student-friendly and technology-savvy formative assessment system
- Provides short, standard-by-standard assessments
- Embedded in classroom practice and available on-demand.
- Provides immediate, actionable results to inform personalized instruction and successfully Navvy-gate each student's learning journey.

5

Slide 11

Video: What are DCMs?

Please go the following video link to watch a short video explaining what DCMs are and how they are different from traditional assessment models.

<https://youtu.be/7fricitlqn8>



What are DCMs?

6

Slide 12

**In-Advance Training for
Standard Setting**

Purpose and Process of Standard Setting

7

Slide 13

The Purpose of the Standard Setting

- The Navvy system produces diagnostic classifications on individual standards for each student.
- Federal policy requires all testing systems produce at least three proficiency designations (i.e., does not meet expectations, meets expectations, exceeds expectations).
 - Georgia uses four proficiency designations (Beginning Learners, Developing Learners, Proficient Learners, Distinguished Learners)
- How do we go from multiple diagnostic classification to the proficiency designations?

8

Slide 14

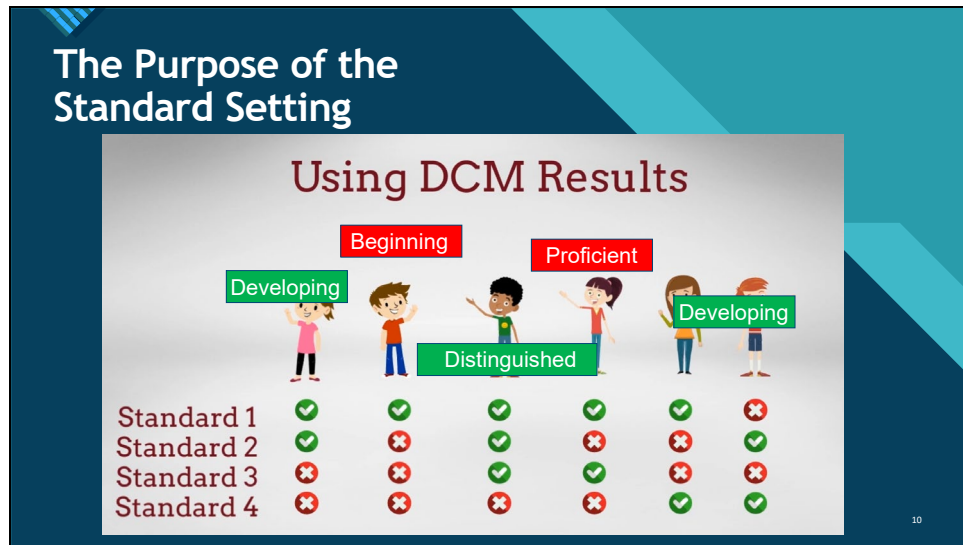
The Purpose of the Standard Setting

Traditional Standardized Test

Math Ability 400-800

Beginning	Developing	Proficient	Distinguished		
425	500	575	650	725	800

9

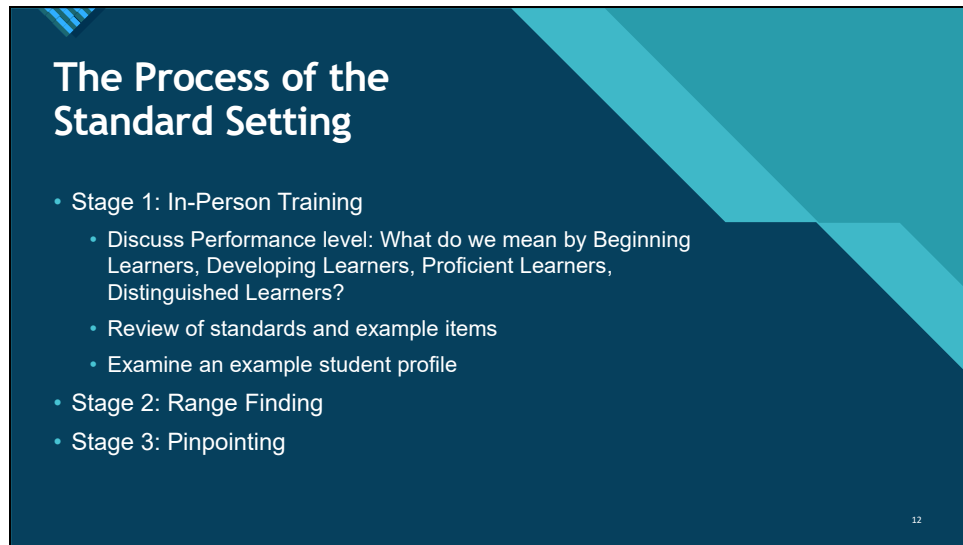


The Process of the Standard Setting

- How do we go from multiple diagnostic classification to the proficiency designations?
- A standard setting procedure that builds consensus among a panel of experts. This process includes three stages:
 - Stage 1: In-Person Training
 - Stage 2: Range Finding
 - Stage 3: Pinpointing

11

Slide 17

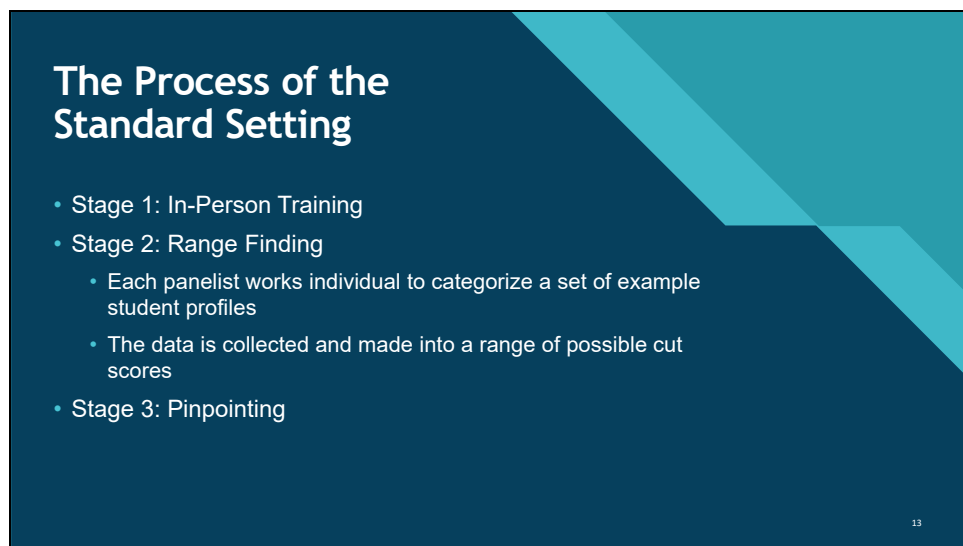


The Process of the Standard Setting

- Stage 1: In-Person Training
 - Discuss Performance level: What do we mean by Beginning Learners, Developing Learners, Proficient Learners, Distinguished Learners?
 - Review of standards and example items
 - Examine an example student profile
- Stage 2: Range Finding
- Stage 3: Pinpointing

12

Slide 18

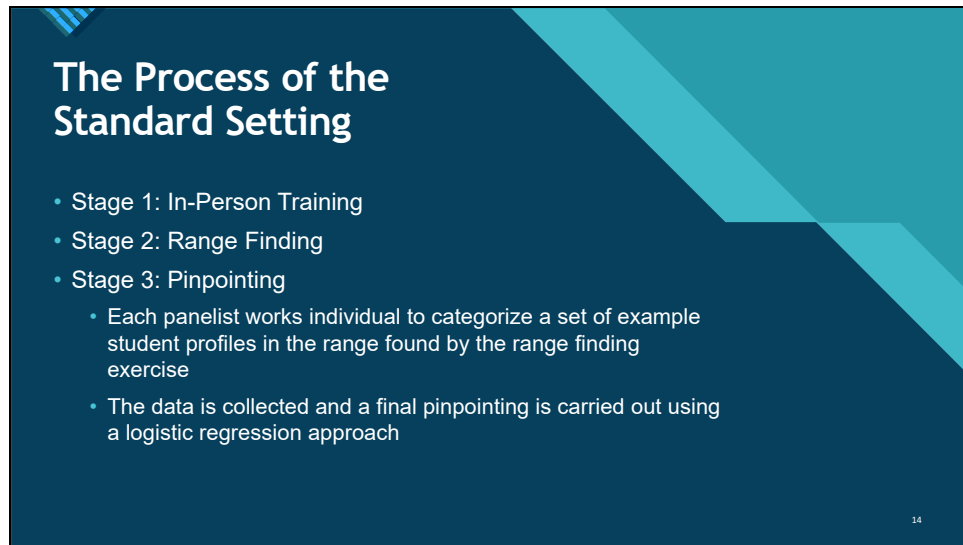


The Process of the Standard Setting

- Stage 1: In-Person Training
- Stage 2: Range Finding
 - Each panelist works individual to categorize a set of example student profiles
 - The data is collected and made into a range of possible cut scores
- Stage 3: Pinpointing

13

Slide 19



The Process of the Standard Setting

- Stage 1: In-Person Training
- Stage 2: Range Finding
- Stage 3: Pinpointing
 - Each panelist works individual to categorize a set of example student profiles in the range found by the range finding exercise
 - The data is collected and a final pinpointing is carried out using a logistic regression approach

14

Slide 20

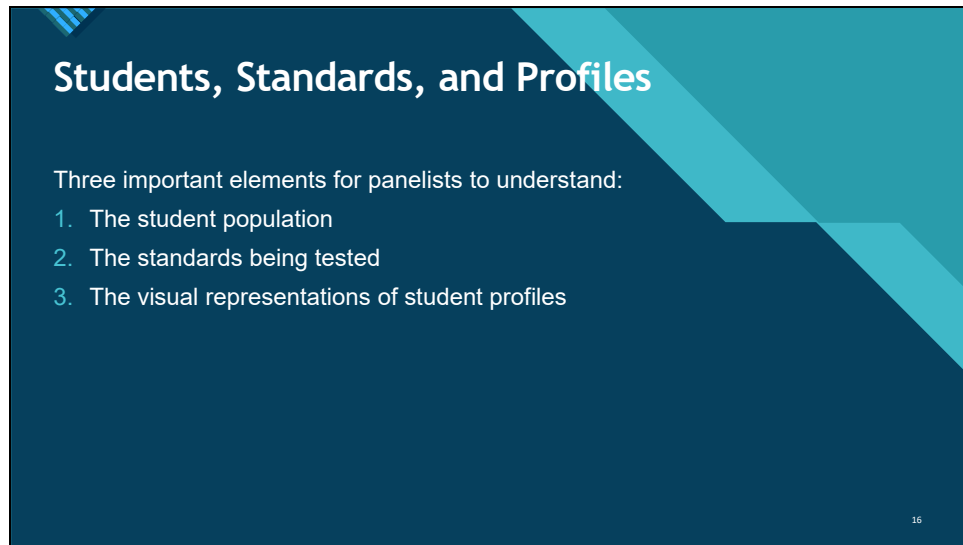


In-Advance Training for Standard Setting

Students, Standards, and Profiles

15

Slide 21

The slide features a dark blue background with a decorative geometric pattern of lighter blue and teal shapes in the top right corner. The title "Students, Standards, and Profiles" is written in white. Below it, a line of text states "Three important elements for panelists to understand:", followed by a numbered list of three items.

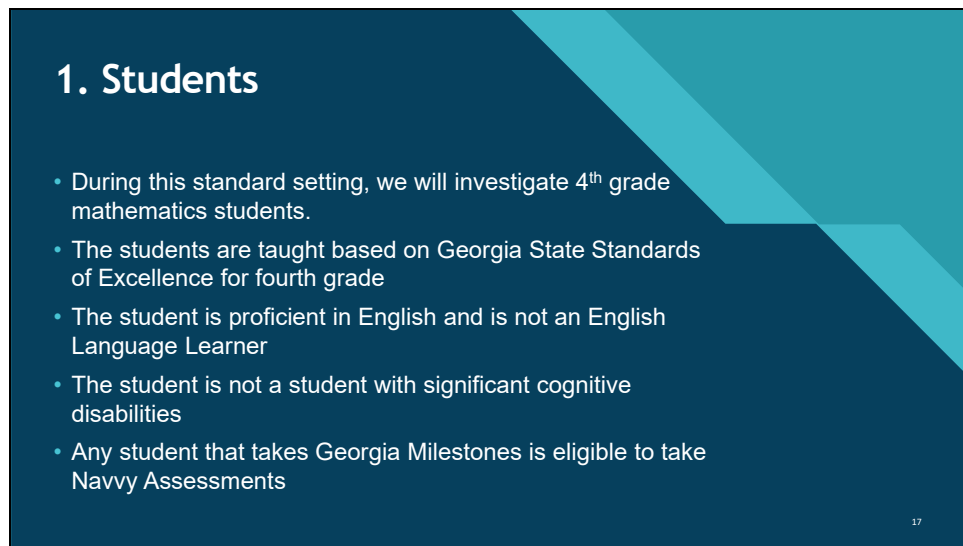
Students, Standards, and Profiles

Three important elements for panelists to understand:

1. The student population
2. The standards being tested
3. The visual representations of student profiles

16

Slide 22

The slide features a dark blue background with a decorative geometric pattern of lighter blue and teal shapes in the top right corner. The title "1. Students" is written in white. Below it, a bulleted list of five criteria is provided.

1. Students

- During this standard setting, we will investigate 4th grade mathematics students.
- The students are taught based on Georgia State Standards of Excellence for fourth grade
- The student is proficient in English and is not an English Language Learner
- The student is not a student with significant cognitive disabilities
- Any student that takes Georgia Milestones is eligible to take Navvy Assessments

17

Slide 23

2. Standards

The general description:

Mathematics | Grade 4

In Grade 4, instructional time should focus on three critical areas: (1) developing understanding and fluency with multi-digit multiplication, and developing understanding of dividing to find quotients involving multi-digit dividends; (2) developing an understanding of fraction equivalence, addition and subtraction of fractions with like denominators, and multiplication of fractions by whole numbers; (3) understanding that geometric figures can be analyzed and classified based on their properties, such as having parallel sides, perpendicular sides, particular angle measures, and symmetry

18

Slide 24

2. Standards

Operations and Algebraic Thinking

- Use the four operations with whole numbers to solve problems.
- Gain familiarity with factors and multiples.
- Generate and analyze patterns.

Number and Operations in Base Ten

- Generalize place value understanding for multi-digit whole numbers.
- Use place value understanding and properties of operations to perform multi-digit arithmetic.

Number and Operations—Fractions

- Extend understanding of fraction equivalence and ordering.
- Build fractions from unit fractions by applying and extending previous understandings of operations on whole numbers.
- Understand decimal notation for fractions, and compare decimal fractions.

Measurement and Data

- Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit.
- Represent and interpret data.
- Geometric measurement: understand concepts of angle and measure angles.

Geometry

- Draw and identify lines and angles, and classify shapes by properties of their lines and angles.

19

3. Profiles

- A “Profile” or a “Student Profile” is a visual representation of the standards a student has or has not mastered.
- You will receive a PDF with a student profile on each page.
- The next several slides show example profiles. Important elements include
 - The standard number, label, and wording
 - The profile label
 - Master/Nonmaster designation by color

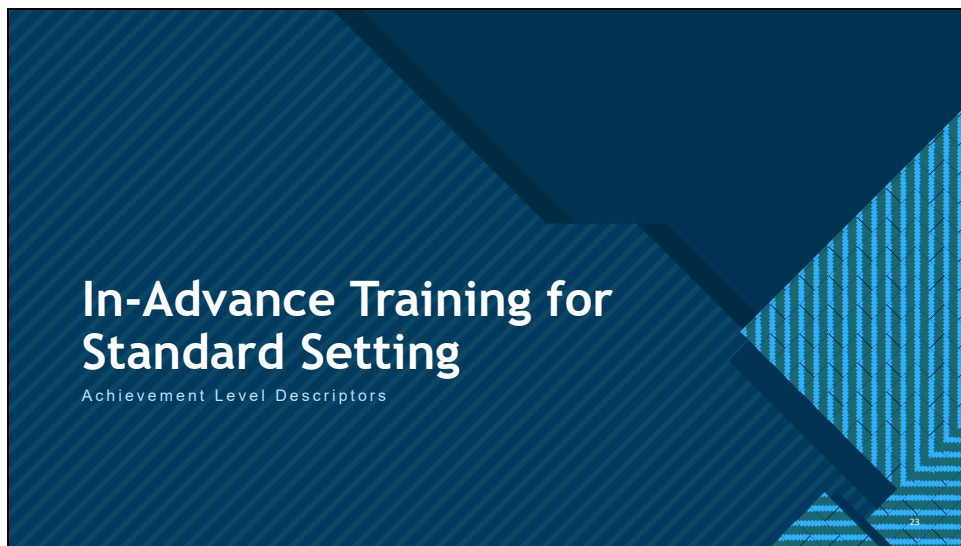
20

Profile Label				
OA = X/5 NBT = X/6 NF = X/7 MD = X/8 G = X/3	1. 4.OA.1 Understand that a multiplicative comparison is a situation in which one quantity is multiplied by a specified number to get another quantity.	2. 4.OA.2 Multiply or divide to solve word problems involving multiplicative comparison. Use drawings and equations with a symbol or letter for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison.	3. 4.OA.3 Solve multistep word problems with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a symbol or letter standing for the unknown quantity. Assess the reasonableness of answers using mental computation and estimation strategies including rounding.	4. 4.OA.4 Find all factor pairs for a whole number in the range 1–100. Recognize that a whole number is a multiple of each of its factors. Determine whether a given whole number in the range 1–100 is a multiple of a given one-digit number. Determine whether a given whole number in the range 1–100 is prime or composite.
	5. 4.OA.5 Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself. Explain informally why the pattern will continue to develop in this way.	6. 4.NBT.1 Recognize that in a multi-digit whole number, a digit in any one place represents ten times what it represents in the place to its right.	7. 4.NBT.2 Read and write multi-digit whole numbers using base-ten numerals, number names, and expanded form. Compare two multi-digit numbers based on meanings of the digits in each place, using >, =, and < symbols to record the results of comparisons.	8. 4.NBT.3 Use place value understanding to round multi-digit whole numbers to any place.
	10. 4.NBT.5 Multiply a whole number of up to four digits by a one-digit whole number, and multiply two-digit numbers, using strategies based on place value and the properties of operations. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	11. 4.NBT.6 Find whole-number quotients and remainders with up to four-digit dividends and one-digit divisors, using strategies based on place value, the properties of operations, and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	12. 4.NF.1 Explain why two or more fractions are equivalent: $\frac{2}{3} = \frac{2 \times 2}{3 \times 2} = \frac{4}{6}$; $\frac{1}{2} = \frac{1 \times 5}{2 \times 5} = \frac{5}{10}$ by using visual fraction models. Focus attention on how the number and size of the parts differ even though the fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions.	13. 4.NF.2 Compare two fractions with different numerators and different denominators, e.g., by using visual fraction models, by creating common denominators or numerators, or by comparing to a benchmark fraction such as $\frac{1}{2}$. Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols >, =, or <, and justify the conclusions.
	15. 4.NF.4 Apply and extend previous understandings of multiplication to multiply a fraction by a whole number e.g., by using a visual such as a number line or area model.	16. 4.NF.5 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	17. 4.NF.6 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	18. 4.NF.7 Compare two decimals to hundredths by reasoning about their size. Recognize that comparisons are valid only when the two decimals refer to the same whole. Record the results of comparisons with the symbols >, =, or <, and justify the conclusions, e.g., by using a visual model.
	20. 4.MD.1 Use the four operations to solve word problems involving distances, intervals of time, liquid volumes, masses of objects, and money, including problems involving simple fractions or decimals, and problems that require expressing measurements given in a larger unit in terms of a smaller unit. Represent measurement quantities using diagrams such as number line diagrams that feature a measurement scale.	21. 4.MD.3 Apply the area and perimeter formulas for rectangles in real world and mathematical problems. For example, find the width of a rectangular room given the area of the flooring and the length, by viewing the area formula as a multiplication equation with an unknown factor.	22. 4.MD.4 Make a line plot to display a data set of measurements in fractions of a unit ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{3}{4}$). Solve problems involving addition and subtraction of fractions with common denominators by using information presented in line plots.	23. 4.MD.5 Recognize angles as geometric shapes that are formed wherever two rays share a common endpoint, and understand concepts of angle measurement.
25. 4.MD.7 Recognize angle measure as additive. When an angle is decomposed into non-overlapping parts, the angle measure of the whole is the sum of the	26. 4.MD.8 Recognize area as additive. Find areas of rectilinear figures by decomposing them into non-overlapping	27. 4.G.1 Draw points, lines, line segments, rays, angles (right, acute, obtuse), and perpendicular and parallel	28. 4.G.2 Classify two-dimensional figures based on the presence or absence of parallel or perpendicular lines, or the presence or absence of	29. 4.G.3 Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded

Slide 27

Profile A14				
OA = 3/5 NBT = 3/6 NF = 3/7 MD = 3/8 G = 2/3	1. 4.OA.1 Understand that a multiplicative comparison is a situation in which one quantity is multiplied by a specified number to get another quantity.	2. 4.OA.2 Multiply or divide to solve word problems involving multiplicative comparison. Use drawings and equations with a symbol or letter for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison.	3. 4.OA.3 Solve multistep word problems with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a symbol or letter standing for the unknown quantity. Assess the reasonableness of answers using mental computation and estimation strategies including rounding.	4. 4.OA.4 Find all factor pairs for a whole number in the range 1–100. Recognize that a whole number is a multiple of each of its factors. Determine whether a given whole number in the range 1–100 is a multiple of a given one-digit number. Determine whether a given whole number in the range 1–100 is prime or composite.
	5. 4.OA.5 Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself. Explain informally why the pattern will continue to develop in this way.	6. 4.NBT.1 Recognize that in a multi-digit whole number, a digit in any one place represents ten times what it represents in the place to its right.	7. 4.NBT.2 Read and write multi-digit whole numbers using base-ten numerals, number names, and expanded form. Compare two multi-digit numbers based on meanings of the digits in each place, using $>$, $=$, and $<$ symbols to record the results of comparisons.	8. 4.NBT.3 Use place value understanding to round multi-digit whole numbers to any place.
	10. 4.NBT.5 Multiply a whole number of up to four digits by a one-digit whole number, and multiply two two-digit numbers, using strategies based on place value and the properties of operations. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	11. 4.NBT.6 Find whole-number quotients and remainders with up to four-digit dividends and one-digit divisors, using strategies based on place value, the properties of operations, and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	12. 4.NF.1 Explain why two or more fractions are equivalent $\frac{1}{2} = \frac{2}{4}$; ex. $\frac{1}{2} = \frac{2 \times 2}{2 \times 4}$ by using visual fraction models. Focus attention on how the number and size of the parts differ even though the fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions.	13. 4.NF.2 Compare two fractions with different numerators and different denominators, e.g., by using visual fraction models, by creating common denominators or numerators, or by comparing to a benchmark fraction such as $\frac{1}{2}$. Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols $>$, $=$, or $<$, and justify the conclusions.
	15. 4.NF.4 Apply and extend previous understandings of multiplication to multiply a fraction by a whole number e.g., by using a visual such as a number line or area model.	16. 4.NF.5 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	17. 4.NF.6 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	14. 4.NF.3 Understand a fraction $\frac{a}{b}$ with a numerator > 1 as a sum of unit fractions $\frac{1}{b}$.
	20. 4.MD.1 Use the four operations to solve word problems involving distances, intervals of time, liquid volumes, masses of objects, and money, including problems that require expressing measurements given in a larger unit in terms of a smaller unit. Represent measurement quantities using diagrams such as number line diagrams that feature a measurement scale.	21. 4.MD.3 Apply the area and perimeter formulas for rectangles in real world and mathematical problems. For example, find the width of a rectangular room given the area of the flooring and the length, by viewing the area formula as a multiplication equation with an unknown factor.	22. 4.MD.4 Make a line plot to display a data set of measurements in fractions of a unit ($\frac{1}{2}$, $\frac{1}{4}$). Solve problems involving addition and subtraction of fractions with common denominators by using information presented in line plots.	18. 4.NF.7 Compare two decimals to hundredths by reasoning about their size. Recognize that comparisons are valid only when the two decimals refer to the same whole. Record the results of comparisons with the symbols $>$, $=$, or $<$, and justify the conclusions, e.g., by using a visual model.
25. 4.MD.7 Recognize angle measure as additive. When an angle is decomposed into non-overlapping parts, the angle measure of the whole is the sum of the	26. 4.MD.8 Recognize area as additive. Find areas of rectilinear figures by decomposing them into non-overlapping	27. 4.G.1 Draw points, lines, line segments, rays, angles (right, acute, obtuse), and perpendicular and parallel	23. 4.MD.5 Recognize angles as geometric shapes that are formed wherever two rays share a common endpoint, and understand concepts of angle measurement.	19. 4.MD.1 Know relative sizes of measurement units within one system of units including km, m, cm; kg, g; lb, oz.; l, ml; hr, min, sec.
			28. 4.G.2 Classify two-dimensional figures based on the presence or absence of parallel or perpendicular lines, or the presence or absence of	24. 4.MD.6 Measure angles in whole-number degrees using a protractor. Sketch angles of specified measure.
			29. 4.G.3 Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded	

Slide 28



Achievement Level Descriptors (ALDs)

- During the standard setting, you will view student profiles and group them as Beginning Learner, Developing Learner, Proficient Learner, and Distinguished Learner.
- But what are each of these learners?
- **Achievement Level Descriptors (ALDs)** describe the learners at different levels of proficiency.
- The next slide provides the policy ALDs.

24

Achievement Level Descriptors (ALDs)

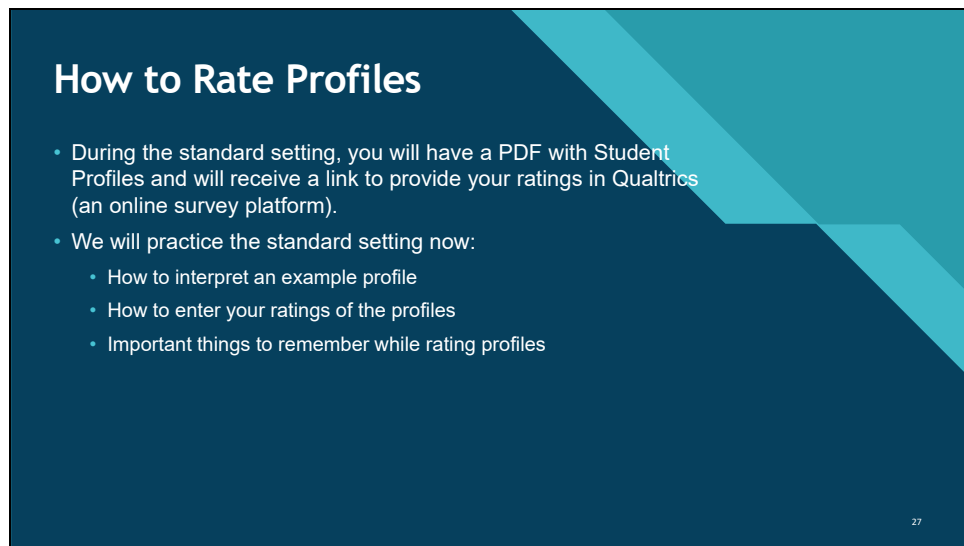
ALD	Standard	Beginning Learner	Developing Learner	Proficient Learner	Distinguished Learner
Policy		Beginning Learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students need substantial academic support to be prepared for the next grade level or course and to be on track for <i>college and career readiness</i> .	Developing Learners demonstrate partial proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students need additional academic support to ensure success in the next grade level or course and to be on track for <i>college and career readiness</i> .	Proficient Learners demonstrate proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students are prepared for the next grade level or course and are on track for <i>college and career readiness</i> .	Distinguished Learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level/course of learning, as specified in Georgia's content standards. The students are well prepared for the next grade level or course and are well prepared for <i>college and career readiness</i> .
Range		A student who achieves at the Beginning Learner level demonstrates minimal command of the grade-level standards.	A student who achieves at the Developing Learner level demonstrates partial command of the grade-level standards.	A student who achieves at the Proficient Learner level demonstrates proficiency of the grade-level standards.	A student who achieves at the Distinguished Learner level demonstrates advanced proficiency of the grade-level standards.

25

Slide 31



Slide 32



Slide 33

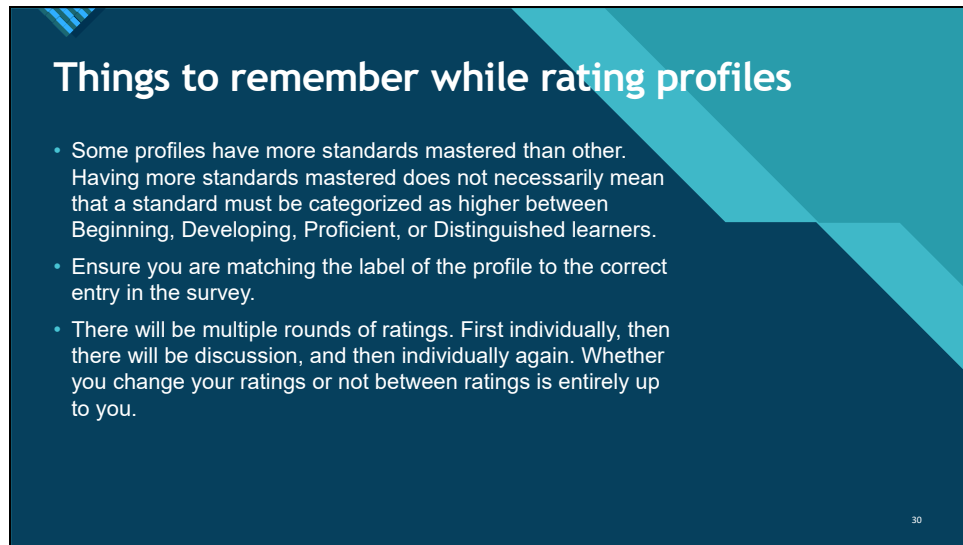
Profile A14				
OA = 3/5 NBT = 3/6 NF = 3/7 MD = 3/8 G = 2/3	1. 4.OA.1 Understand that a multiplicative comparison is a situation in which one quantity is multiplied by a specified number to get another quantity.	2. 4.OA.2 Multiply or divide to solve word problems involving multiplicative comparison. Use drawings and equations with a symbol or letter for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison.	3. 4.OA.3 Solve multistep word problems with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a symbol or letter standing for the unknown quantity. Assess the reasonableness of answers using mental computation and estimation strategies including rounding.	4. 4.OA.4 Find all factor pairs for a whole number in the range 1–100. Recognize that a whole number is a multiple of each of its factors. Determine whether a given whole number in the range 1–100 is a multiple of a given one-digit number. Determine whether a given whole number in the range 1–100 is prime or composite.
	5. 4.OA.5 Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself. Explain informally why the pattern will continue to develop in this way.	6. 4.NBT.1 Recognize that in a multi-digit whole number, a digit in any one place represents ten times what it represents in the place to its right.	7. 4.NBT.2 Read and write multi-digit whole numbers using base-ten numerals, number names, and expanded form. Compare two multi-digit numbers based on meanings of the digits in each place, using $>$, $=$, and $<$ symbols to record the results of comparisons.	8. 4.NBT.3 Use place value understanding to round multi-digit whole numbers to any place.
	10. 4.NBT.5 Multiply a whole number of up to four digits by a one-digit whole number, and multiply two two-digit numbers, using strategies based on place value and the properties of operations. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	11. 4.NBT.6 Find whole-number quotients and remainders with up to four-digit dividends and one-digit divisors, using strategies based on place value, the properties of operations, and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	12. 4.NF.1 Explain why two or more fractions are equivalent $\frac{1}{2} = \frac{2}{4}$; ex. $\frac{1}{2} = \frac{2}{4}$ by using visual fraction models. Focus attention on how the number and size of the parts differ even though the fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions.	13. 4.NF.2 Compare two fractions with different numerators and different denominators, e.g., by using visual fraction models, by creating common denominators or numerators, or by comparing to a benchmark fraction such as $\frac{1}{2}$. Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols $>$, $=$, or $<$, and justify the conclusions.
	15. 4.NF.4 Apply and extend previous understandings of multiplication to multiply a fraction by a whole number e.g., by using a visual such as a number line or area model.	16. 4.NF.5 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	17. 4.NF.6 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	18. 4.NF.7 Compare two decimals to hundredths by reasoning about their size. Recognize that comparisons are valid only when the two decimals refer to the same whole. Record the results of comparisons with the symbols $>$, $=$, or $<$, and justify the conclusions, e.g., by using a visual model.
	20. 4.MD.1 Use the four operations to solve word problems involving distances, intervals of time, liquid volumes, masses of objects, and money, including problems that require expressing measurements given in a larger unit in terms of a smaller unit. Represent measurement quantities using diagrams such as number line diagrams that feature a measurement scale.	21. 4.MD.3 Apply the area and perimeter formulas for rectangles in real world and mathematical problems. For example, find the width of a rectangular room given the area of the flooring and the length, by viewing the area formula as a multiplication equation with an unknown factor.	22. 4.MD.4 Make a line plot to display a data set of measurements in fractions of a unit ($\frac{1}{2}$, $\frac{1}{4}$). Solve problems involving addition and subtraction of fractions with common denominators by using information presented in line plots.	23. 4.MD.5 Recognize angles as geometric shapes that are formed wherever two rays share a common endpoint, and understand concepts of angle measurement.
25. 4.MD.7 Recognize angle measure as additive. When an angle is decomposed into non-overlapping parts, the angle measure of the whole is the sum of the	26. 4.MD.8 Recognize area as additive. Find areas of rectilinear figures by decomposing them into non-overlapping	27. 4.G.1 Draw points, lines, line segments, rays, angles (right, acute, obtuse), and perpendicular and parallel	28. 4.G.2 Classify two-dimensional figures based on the presence or absence of parallel or perpendicular lines, or the presence or absence of	19. 4.MD.1 Know relative sizes of measurement units within one system of units including km, m, cm; kg, g; lb, oz.; l, ml; hr, min, sec.
				24. 4.MD.6 Measure angles in whole-number degrees using a protractor. Sketch angles of specified measure.
				29. 4.G.3 Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded

Slide 34

How to interpret an example profile

- During the standard setting, you will have a PDF with Student Profiles and will receive a link to provide your ratings in Qualtrics (an online survey platform).
- We will practice the standard setting now:
 - How to interpret an example profile
 - How to enter your ratings of the profiles
 - Important things to think about while rating profiles

Slide 35



Things to remember while rating profiles

- Some profiles have more standards mastered than other. Having more standards mastered does not necessarily mean that a standard must be categorized as higher between Beginning, Developing, Proficient, or Distinguished learners.
- Ensure you are matching the label of the profile to the correct entry in the survey.
- There will be multiple rounds of ratings. First individually, then there will be discussion, and then individually again. Whether you change your ratings or not between ratings is entirely up to you.

30

Slide 36



In-Advance Training for Standard Setting

Survey and Self-Assessment

31

In-Advance Training Survey

The following survey is meant to be a self-assessment.

- Your answers are meant to get an understanding of standard setting participants (demographics and experiences) and to see if there are areas or topics that need to be discussed further during the in-person standard setting.
- Your answers will not determine or change your eligibility to participate. It is only important to answer as honestly as possible.

Please following the following link to participate in the survey

https://ugeorgia.ca1.qualtrics.com/jfe/form/SV_b75MTDRLDkJNNyu

Appendix I

IRB Approval Letter



Tucker Hall, Room 212
310 E. Campus Rd.
Athens, Georgia 30602
TEL 706-542-3199 | FAX 706-542-5638
IRB@uga.edu
<http://research.uga.edu/hso/irb/>

Human Research Protection Program

NOT HUMAN RESEARCH DETERMINATION

June 10, 2022

Dear [Laine Bradshaw](#):

On 6/10/2022, the Human Subjects Office reviewed the following submission:

Title of Study:	Profile Selection, Variability, and Range in Standard Setting for Diagnostic Classification Models
Investigator:	Laine Bradshaw
Co-Investigator:	Zachary Feldberg
IRB ID:	PROJECT00005886
Funding:	None

We have determined that the proposed activity is not designed as research involving human subjects as defined by DHHS and FDA regulations. The activity is designed to develop variables in diagnostic classification models. The investigators are focused on getting expert feedback on development of standards rather than the experts.

University of Georgia (UGA) IRB review and approval is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human subjects, please submit a new request to the IRB for a determination.

Sincerely,

Jessica Lasebikan, HRPP Assistant Director
Human Subjects Office, University of Georgia

Appendix J

Profile Card

<p>Profile A14</p> <p>NBT = 4/6 G = 2/3 OA = 2/5 NF = 1/ 7 MD = 5/ 8</p>	<p>1. 4.NBT.1 Recognize that in a multi-digit whole number, a digit in any one place represents ten times what it represents in the place to its right.</p>	<p>2. 4.NBT.2 Read and write multi-digit whole numbers using base-ten numerals, number names, and expanded form. Compare two multi-digit numbers based on meanings of the digits in each place, using $>$, $=$, and $<$ symbols to record the results of comparisons.</p>	<p>3. 4.NBT.3 Use place value understanding to round multi-digit whole numbers to any place.</p>
<p>7. 4.G.1 Draw points, lines, line segments, rays, angles (right, acute, obtuse), and perpendicular and parallel lines. Identify these in two-dimensional figures.</p>	<p>8. 4.G.2 Classify two-dimensional figures based on the presence or absence of parallel or perpendicular lines, or the presence or absence of angles of a specified size. Recognize right triangles as a category, and identify right triangles.</p>	<p>9. 4.G.3 Recognize a line of symmetry for a two-dimensional figure as a line across the figure such that the figure can be folded along the line into matching parts. Identify line-symmetric figures and draw lines of symmetry.</p>	<p>10. 4.OA.1 Understand that a multiplicative comparison is a situation in which one quantity is multiplied by a specified number to get another quantity.</p>
<p>15. 4.NF.1 Explain why two or more fractions are equivalent $\frac{a}{b} = \frac{n \times a}{n \times b}$ ex: $\frac{1}{4} = \frac{3 \times 1}{3 \times 4}$ by using visual fraction models. Focus attention on how the number and size of the parts differ even though the fractions themselves are the same size. Use this principle to recognize and generate equivalent fractions.</p>	<p>16. 4.NF.2 Compare two fractions with different numerators and different denominators, e.g., by using visual fraction models, by creating common denominators or numerators, or by comparing to a benchmark fraction such as $\frac{1}{2}$. Recognize that comparisons are valid only when the two fractions refer to the same whole. Record the results of comparisons with symbols $>$, $=$, or $<$, and justify the conclusions.</p>	<p>17. 4.NF.3 Understand a fraction $\frac{a}{b}$ with a numerator > 1 as a sum of unit fractions $\frac{1}{b}$.</p>	<p>18. 4.NF.4 Apply and extend previous understandings of multiplication to multiply a fraction by a whole number e.g., by using a visual such as a number line or area model.</p>
<p>22. 4.MD.1 Know relative sizes of measurement units within one system of units including km, m, cm; kg, g; lb, oz.; l, ml; hr, min, sec.</p>	<p>23. 4.MD.1 Use the four operations to solve word problems involving distances, intervals of time, liquid volumes, masses of objects, and money, including problems involving simple fractions or decimals, and problems that require expressing measurements given in a larger unit in terms of a smaller unit. Represent measurement quantities using diagrams such as number line diagrams that feature a measurement scale.</p>	<p>24. 4.MD.3 Apply the area and perimeter formulas for rectangles in real world and mathematical problems. For example, find the width of a rectangular room given the area of the flooring and the length, by viewing the area formula as a multiplication equation with an unknown factor.</p>	<p>25. 4.MD.4 Make a line plot to display a data set of measurements in fractions of a unit ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$). Solve problems involving addition and subtraction of fractions with common denominators by using information presented in line plots.</p>

Note. The first half of the profile card. The second half continues to the right.

4. 4.NBT.4 Fluently add and subtract multi-digit whole numbers using the standard algorithm.	5. 4.NBT.5 Multiply a whole number of up to four digits by a one-digit whole number, and multiply two two-digit numbers, using strategies based on place value and the properties of operations. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	6. 4.NBT.6 Find whole-number quotients and remainders with up to four-digit dividends and one-digit divisors, using strategies based on place value, the properties of operations, and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.	
11. 4.OA.2 Multiply or divide to solve word problems involving multiplicative comparison. Use drawings and equations with a symbol or letter for the unknown number to represent the problem, distinguishing multiplicative comparison from additive comparison.	12. 4.OA.3 Solve multistep word problems with whole numbers and having whole-number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a symbol or letter standing for the unknown quantity. Assess the reasonableness of answers using mental computation and estimation strategies including rounding.	13. 4.OA.4 Find all factor pairs for a whole number in the range 1–100. Recognize that a whole number is a multiple of each of its factors. Determine whether a given whole number in the range 1–100 is a multiple of a given one-digit number. Determine whether a given whole number in the range 1–100 is prime or composite.	14. 4.OA.5 Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself. Explain informally why the pattern will continue to develop in this way.
19. 4.NF.5 Express a fraction with denominator 10 as an equivalent fraction with denominator 100, and use this technique to add two fractions with respective denominators 10 and 100.	20. 4.NF.6 Use decimal notation for fractions with denominators 10 or 100.	21. 4.NF.7 Compare two decimals to hundredths by reasoning about their size. Recognize that comparisons are valid only when the two decimals refer to the same whole. Record the results of comparisons with the symbols $>$, $=$, or $<$, and justify the conclusions, e.g., by using a visual model.	
26. 4.MD.5 Recognize angles as geometric shapes that are formed wherever two rays share a common endpoint, and understand concepts of angle measurement.	27. 4.MD.6 Measure angles in whole-number degrees using a protractor. Sketch angles of specified measure.	28. 4.MD.7 Recognize angle measure as additive. When an angle is decomposed into non-overlapping parts, the angle measure of the whole is the sum of the angle measures of the parts. Solve addition and subtraction problems to find unknown angles on a diagram in real world and mathematical problems, e.g., by using an equation with a symbol or letter for the unknown angle measure.	29. 4.MD.8 Recognize area as additive. Find areas of rectilinear figures by decomposing them into non-overlapping rectangles and adding the areas of the non-overlapping parts, applying this technique to solve real world problems.

Note. The second half of the profile card. The first portions precedes from the left.

Appendix K

Links to Packets of Profiles Mailed to Panelists

Condition/ Panelist	Link
Conditions 1, 2, 4, 5	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/ERhglYRSNTpNv64WuRFis-oB2yAGMUkSE2iezyugNMkT3Q
Condition 3 Panelist 1	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/EULN9wdO5UdAnNDDp02USvIBcJvqM9oW7v0MAEauRkzFEg?e=d5soqF
Condition 3 Panelist 2	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/ETtaYgOFJclFhAWJ_ss2GHsBNgNM866BgrPFIgGExAa1Ew?e=xlkv6x
Condition 3 Panelist 3	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/EZZCYs8Ji0NEo-NGXUMYU7cBbgqYNSJVRooypymaUOj5Iw
Condition 3 Panelist 4	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/EXQDE9GC18BBqsw659EmHpYBbnd6ebJwgl1FTykMbaoXCA?e=gilckj
Condition 3 Panelist 5	https://outlookuga-my.sharepoint.com/:b:/g/personal/feldberg_uga_edu/EfXdDvh8s6pHk97YPoW-Oj4B2P5JVF03nQCoLPq9jOJzQg?e=0BrXKy

Appendix L

Example of Data Displayed During Discussion Phases

	B	M	N	O	P	Q	R
1	Profile	Panel1	Panel2	Panel3	Panel4	Panel 5	Variance
2	A03	1	1	1	1	1	0
3	B03	1	1	1	1	1	0
4	C03	1	1	1	1	1	0
5	A06	1	1	1	1	1	0
6	B06	1	1	1	2	1	0.16
7	C06	1	1	1	2	1	0.16
8	A09	1	2	2	2	2	0.16
9	B09	1	1	2	2	2	0.24
10	C09	2	1	2	2	1	0.24
11	A12	2	2	2	2	2	0
12	B12	2	3	2	2	2	0.16
13	C12	2	3	2	2	2	0.16
14	A15	2	3	3	2	2	0.24
15	B15	2	2	2	3	2	0.16
16	C15	2	2	2	3	2	0.16
17	A18	2	3	3	3	3	0.16
18	B18	2	3	3	3	3	0.16
19	C18	2	2	2	3	3	0.24
20	A21	3	3	3	3	4	0.16
21	B21	3	3	3	3	4	0.16
22	C21	3	3	3	3	4	0.16
23	A24	3	4	4	3	3	0.24
24	B24	3	4	4	3	3	0.24
25	C24	4	4	3	3	4	0.24
26	A27	4	4	4	4	4	0
27	B27	4	4	4	4	4	0
28	C27	4	4	4	4	4	0

Note. The profile is the label of the student attribute profile categorized by the judge. The ratings under each panelist represent 1 for Beginning Learner, 2 is for Developing Learner, 3 for Proficient Learner, and 4 is for Distinguished Learner. Variance is calculated across the ratings.

Appendix M

Standard Deviations and Standard Errors of Total Mastery Level by Proficiency Level Rating and Condition

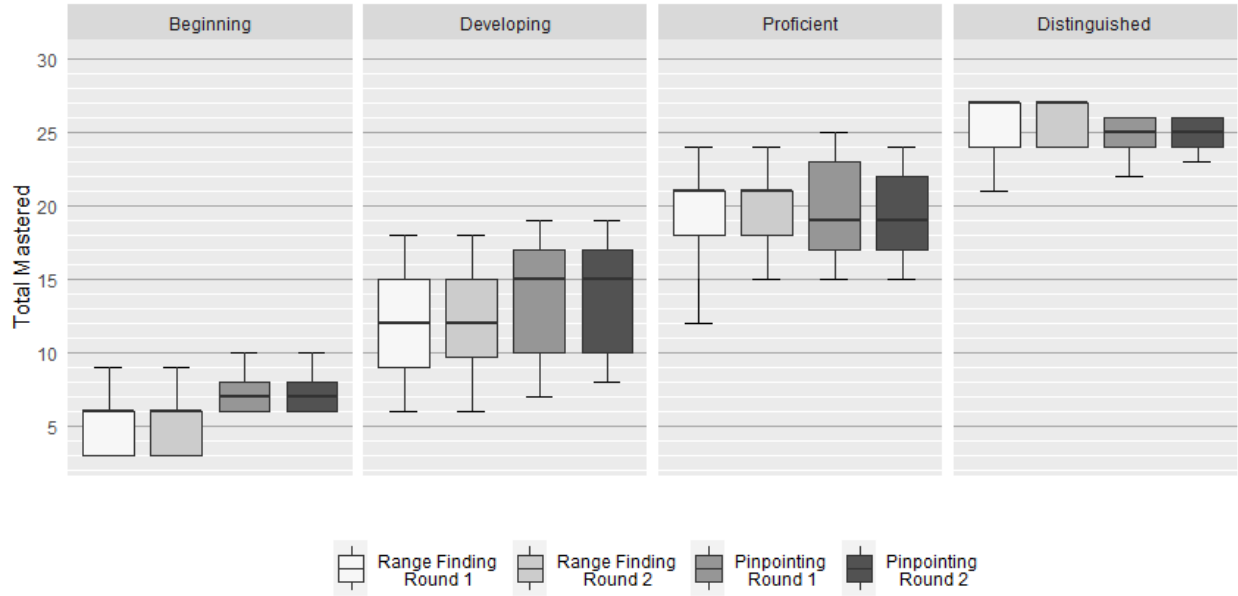
	Beginning	Developing	Proficient	Distinguished
Condition 1				
Range Finding 1	2.21 (0.39)	3.39 (0.52)	3.40 (0.57)	2.14 (0.43)
Range Finding 2	2.08 (0.36)	3.13 (0.46)	2.21 (0.42)	1.52 (0.29)
Pinpointing 1	1.28 (0.19)	3.97 (0.49)	2.96 (0.36)	1.13 (0.17)
Pinpointing 2	1.20 (0.17)	3.72 (0.46)	2.76 (0.34)	0.97 (0.15)
Condition 2				
Range Finding 1	2.35 (0.46)	2.44 (0.52)	2.15 (0.46)	1.54 (0.34)
Range Finding 2	2.29 (0.46)	2.28 (0.47)	1.79 (0.39)	1.54 (0.34)
Pinpointing 1	1.32 (0.25)	2.53 (0.34)	2.74 (0.39)	1.03 (0.27)
Pinpointing 2	1.32 (0.25)	2.53 (0.34)	2.55 (0.38)	0.99 (0.22)
Condition 3				
Range Finding 1	2.34 (0.4)	3.94 (0.59)	3.86 (0.63)	2.04 (0.46)
Range Finding 2	2.34 (0.4)	3.79 (0.58)	3.81 (0.62)	2.04 (0.46)
Pinpointing 1	1.44 (0.24)	4.17 (0.48)	3.61 (0.42)	1.37 (0.22)
Pinpointing 2	1.44 (0.24)	4.16 (0.47)	3.61 (0.42)	1.37 (0.22)
Condition 4				
Range Finding 1	3.95 (0.76)	3.77 (0.56)	3.55 (0.62)	NA
Range Finding 2	3.39 (0.68)	3.84 (0.58)	3.17 (0.53)	NA
Pinpointing 1	1.52 (0.27)	4.65 (0.59)	1.44 (0.19)	NA
Pinpointing 2	1.48 (0.25)	4.56 (0.61)	1.44 (0.19)	NA
Condition 5				
Range Finding 1	NA	2.64(0.52)	4.00 (0.69)	2.79 (0.57)
Range Finding 2	NA	2.63(0.5)	3.86 (0.67)	2.47 (0.51)
Pinpointing 1	NA	1.38(0.25)	3.07 (0.38)	1.31 (0.26)
Pinpointing 2	NA	1.43(0.26)	3.04 (0.38)	1.33 (0.26)

Note. Number outside of parenthesis is the standard deviation. Number inside the parenthesis is the standard error. Condition 4 and Condition 5 did not categorize Distinguished or Beginning Learners, respectfully, as indicated by the “NA” for Not Applicable.

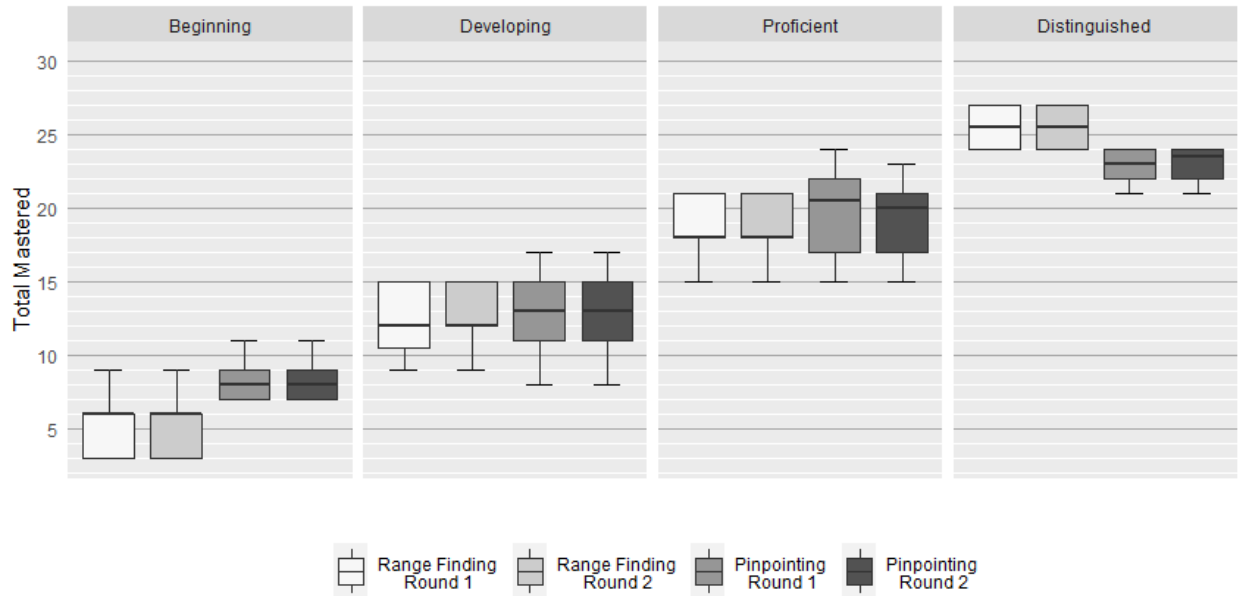
Appendix N

Convergence Plots

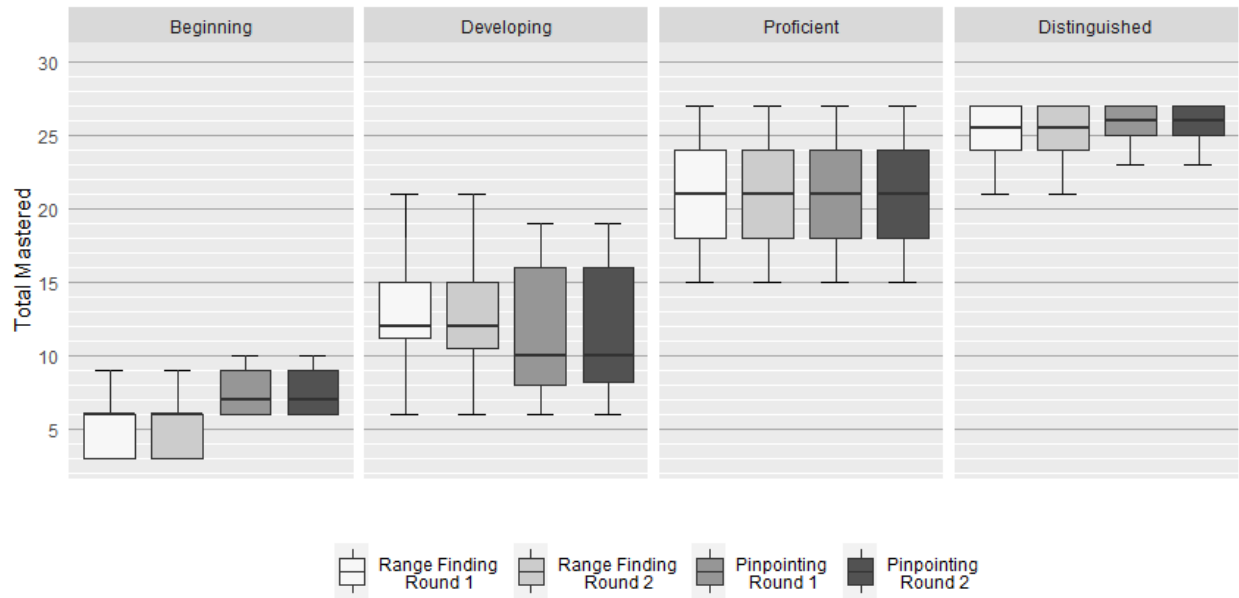
Convergence Plot of Condition 1



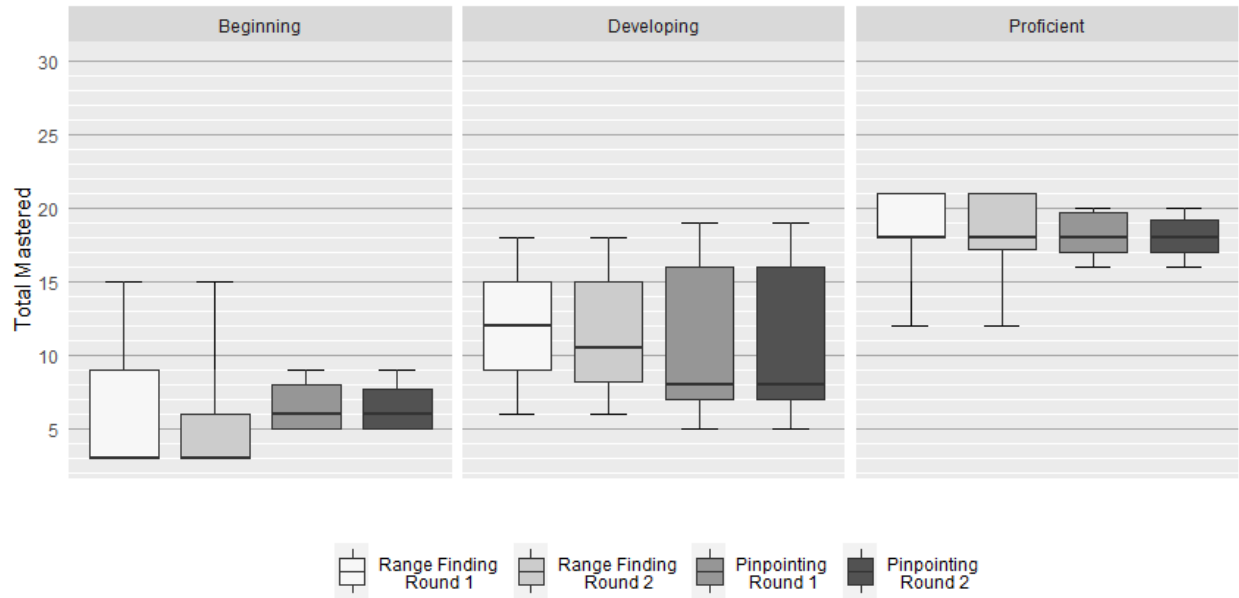
Convergence Plot of Condition 2



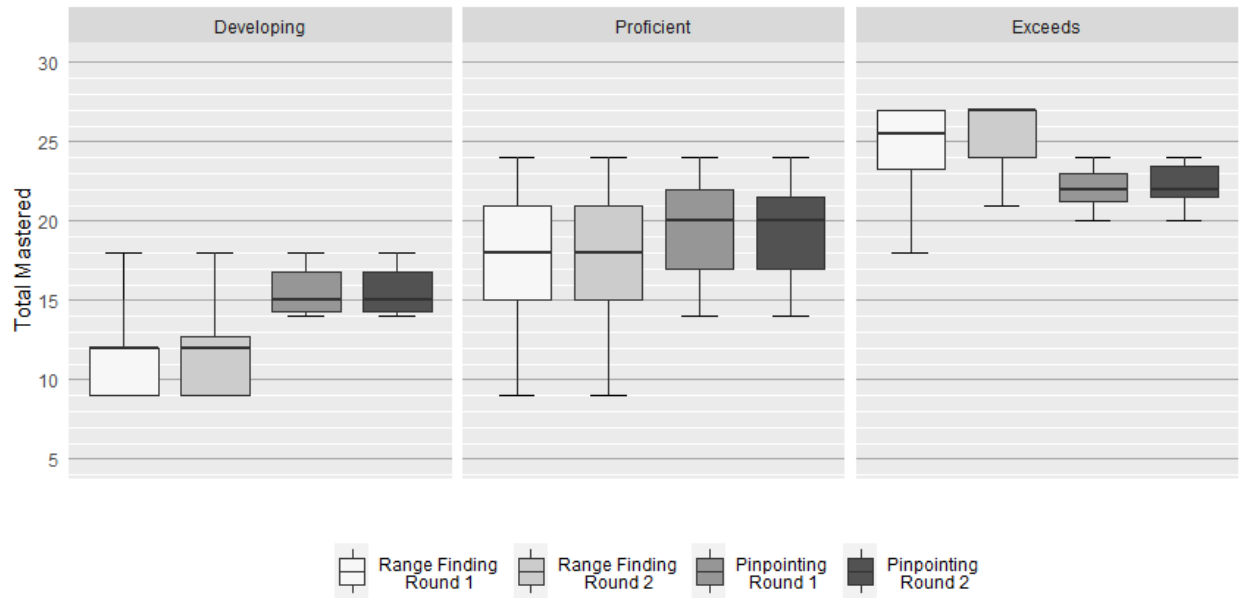
Convergence Plot of Condition 3



Convergence Plot of Condition 4



Convergence Plot of Condition 5



Appendix O

Panelist Exit Survey Results

	Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree	Not Applicable
The goals of the standard setting were met	20	3	0	1	0
The In-advance training prepared me for the standard setting	20	2	0	1	1
The training at the start of the standard setting prepared me for the standard setting	20	2	0	1	1
The standard setting session was well organized	21	1	0	1	1
It was clear what knowledge, skills, or abilities were associated with a certain profile	20	3	0	1	0
I considered the achievement level descriptors when rating each profile	21	2	0	1	0
I considered other panelists opinions when I rated each profile	20	1	2	1	0
I considered my experience when I rated each profile	21	2	0	1	0
I have enough time to complete the required tasks	21	1	0	1	1
I understood how to complete the rating task	22	0	0	1	1
I feel confident in the ratings I made	21	1	0	1	1
I feel confident in the ratings the panel made	20	2	0	1	1
The procedure for recommending cut points was free of bias	20	3	0	1	0
If asked, I would defend the groups cut point for the distinction between "Developing Learners" and "Proficient Learners."	19	3	1	1	0
If asked, I would defend the groups cut point for the distinction between "Beginning Learners" and "Developing Learners."	18	1	0	1	4
If asked, I would defend the groups cut point for the distinction between "Proficient Learners" and "Distinguished Learners."	16	2	0	1	5

Panelist Exit Survey Results (continued)

	Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree	Not Applicable
Participating in the standard setting improved my understanding of assessment	21	3	0	0	0
Participating in the standard setting improved my understanding of the student population	15	6	3	0	0
Participating in the standard setting improved my understanding of Diagnostic Classification Models	18	6	0	0	0
Participating in the standard setting improved my understanding of the Navvy Assessment System	19	5	0	0	0
Overall, I valued my participation in the standard setting	23	0	0	0	1

Appendix P

Panelist Exit Survey Comments

“Great presentation. I enjoyed the perspectives of the other teachers”

“This session provided valuable information about [Diagnostic Assessment Systems] and other tools to guide instruction. It was pleasant hearing the professional opinion and logic of other educators at different levels.”

“The standard setting was a great opportunity for me to have my voice heard and offer my professional insight into how students are classified as learners.

“[The facilitator] did a great job of explaining the expectations and leading us through the data.”

“No comments at this time”

"I think it would have been better if we all had the same profiles so we could compare apples to apples. It helped when you put the other peoples profiles on the screen.”

“This helped me a lot to think about the Georgia Standards and why we rate them the way we do.”

“For my first answer, it made me put in a number however; it depends on which standards they mastered, not just a value. That is just my opinion.”

“I appreciate this experience.”

“I really enjoyed having an open discussion with other teachers about how standards are measured. It was nice to see an up-close picture of each profile.”

“I noticed that as math teachers and leaders in math in Georgia, us as panelists looked at the data as what the students need to know for fourth grade and to do well on the milestones. However, when we looked at the final distinctions the results were based on how many standards the student had mastered, not what standards the student had actually mastered.”

“Thank you for allowing me to participate. I thought it was very informative and a good use of time. It is helpful to always continue the learning process. Understanding what DCM means is helpful in case it comes up in future discussions.”

“This was an excellent session!”

“This was a great opportunity to work with educators beyond my district to assess how to service my students when it comes to math instruction”

“This was really valuable experience. I gain great insight from the panel.”

“Fantastic study!!!!”

“This standard setting was very helpful and insightful. It helped me pin-point the slight differences between proficient learners and distinguished learners. I think having standard settings like this would be beneficial for all teachers in order to better understand what is considered to be developing, proficient, and distinguished.”