Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 461-485

Standard setting in complex performance assessments: An approach aligned with cognitive diagnostic models[1]

Robert W. Lissitz[2] & Feifei Li[3]

Abstract

With the increased interest in student-level diagnostic information from multiple performance assessments, it becomes possible to create multivariate classifications of knowledge, skills and abilities (KSAs). In this paper, a systematic, multivariate and non-compensating standard setting approach, called the cognitive analytical approach (CAA), is proposed for performance assessment with complex tasks.

CAA is based on the framework of evidence-centered design (Mislevy, Steinberg, & Almond, 2003) that supports a chain of reasoning from design and development to delivery of an assessment. In CAA, the performance standards are established simultaneously with domain-modeling, test specifications, and item writing rather than after the assessment has been completed; the cut scores are evaluated iteratively along with the test design and development phases. CAA has the benefits of ensuring the validity of the performance standards, reducing the cognitive load of standard setting, including the complexity of the tasks, and facilitating the vertical articulation of KSAs. In this paper, we elucidate the theoretical and practical rationale of CAA and demonstrate its procedures and results with an illustrative example.

Key words: Standard setting, cognitive diagnostic models, analytical approach, evidence centered design, performance centered, multidimensional standards

[1] An earlier version of this paper was presented at the Annual Meeting of NCME in Denver, Colorado (Lissitz and Li, 2010). The two authors should share equally in any compliments and any criticism that this paper might receive.

[2] *Correspondence concerning this article should be addressed to:* Robert W. Lissitz, PhD, 1229 Benjamin Building, University of Maryland, College Park, MD 20742, USA; email: RLissitz@umd.edu

[3] University of Pennsylvania

Introduction

Setting performance standards is critically important because they are used to determine which examinees will be certified, licensed, or graduated. In the context of No Child Left Behind that is mandated by the Federal Government (NCLB, 2001), individual students' academic achievements are evaluated through state testing. As a result of the evaluation, each student is assigned a Performance Level Label (PLL) based on these performance standards. One example set of PLLs could be "basic", "proficient", and "advanced". Cut scores are intended to divide students into each performance category. These standard-based labels have become an effective means of communicating the results to a variety of audiences, including parents, teachers, administrators and policymakers, and the proportion of proficient or above proficient students in a school/district may be used to determine whether the school is performing satisfactorily over time.

Despite its significance in testing and the educational system, the procedure of standard setting is often seen as arbitrary (Glass, 1978), because little consensus is often reached on the best choice of procedures, and the results of standard setting cannot be easily validated post hoc (Kane, 1994). In addition to producing defensible and valid performance standards by selecting an appropriate method and following the rigorous procedural guidelines, some scholars argue that the results of the standard setting should be evaluated in a validity framework (Hambleton & Pitoniak, 2006; Cizek, 1996). Some of them also suggested that performance standards be set in line with the design model of the assessment so that the tests could be developed on the targeted constructs and created to fit the standard (Bejar, Braun, & Tannenbaum, 2007; Bejar, 2008; Kane, 1994).

In addition to the need for a cognitive framework, there has recently been an increasing interest in the finer-grained student-level diagnostic information from performance assessment (DiBello, Roussos, & Stout, 2007). For example, NCLB requires the parents, teachers and principals receive a diagnostic report to ensure the student obtains the necessary level of knowledge, skills and abilities (KSAs) (Goodman & Hambleton, 2004). The fine-grained diagnostic feedback makes it possible for the individuals, instructors or the program managers to identify the deficiencies in abilities that are revealed by the content standards and implement interventions to remedy those skills that have not yet been mastered.

For the traditional standard setting methods that fall in a test-centered vs. examinee-centered classification (e.g., bookmark, Angoff), a single unidimensional continuum is assumed along which either the difficulty of items or the ability of the examinees can be rank ordered. In contrast, current performance assessments with complex tasks require the tasks be developed based on a well-established cognitive model so as to ensure the link with the KSAs of interest and draw sensible inferences from the scores. For items that involve multiple KSAs, a single continuum or even a composite scale may not capture multiple KSAs that underlie a complex task.

In response, new standard-setting methods for multidimensional tests have been created for educational assessments that include constructed-response items such as writing samples and short answer questions. These new methods either involve the review of candidate work or the review of the score profiles (Hambleton & Pitoniak, 2006). When the panelists are required to select the borderline work or rank order the work based on their quality, standards are set with respect to the overall quality of the examinees' performance across all questions. In contrast, it might be more informative to create classifications for each of the multiple KSAs and profile the examinees. In the standard setting methods involving the review of the score profiles, the standards are presented as score vectors, the purpose of which is to capture multiple KSAs of tests containing complex multidimensional tasks (Jaeger, 1995a, 1995b; Plake, Hambleton, & Jaeger, 1997). Although there is evidence indicating the feasibility and reliability of this type of method, the implementation procedure is challenging as it is not easy to explain the statistical models and the overall process to the panelists.

Some researchers (Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007) proposed using probabilistic diagnostic models to estimate the cut scores and classify the students. This is regarded as an objective standard setting approach in which the classifications are subject to the properties of the items and the performance of the population. However, if the number of examinees is not large enough, the model will be unidentified. In addition, the probabilistic diagnostic models are quite complicated statistical approaches that may not be appropriate for most of the audiences that use the score reports.

The shortcomings of each of the approaches above have limited their contribution to the standard-setting for complex performance assessment. According to Hambleton and his colleagues (Hambleton, Jaeger, Plake, & Mills, 2000), standard-setting for performance assessment is not nearly as well developed, and none of the methods have been fully researched and validated. Standard 4.21 in the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999) states that "When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way." (4.21, p 60) It stresses the importance of designing a process where panelists can optimally use the knowledge that they have to influence the process.

The purpose of this article is to propose a systematic, multivariate and non-compensating (i.e., one higher skill does not compensate for another lower skill) standard setting approach for performance assessment with complex tasks, termed the "Cognitive Analytical Approach" (CAA). CAA is created based on the framework of evidence-centered design (Mislevy, et al., 2003; Kane, 2004). In CAA, the performance standards are established simultaneously with domain modeling, test specifications and item writing; the cut scores are evaluated iteratively before and after the test development phases. By using this procedure, we expect to ensure the validity of the tests and performance standards, reduce the cognitive complexity of standard setting, and facilitate the vertical articulation of KSAs. In this paper, we intend to answer the following questions 1) What is the theoretical rationale for the CAA approach? 2) Why might the CAA be appropriate for standard setting in cognitive diagnostic assessments, compared with other approaches? 3) How should the CAA result be presented and 4) How should CAA results be used?

To address these questions, we first briefly illustrate the theoretical components for this standard setting approach, including the theories of cognitive diagnostic assessment design. Next, we make an argument for the hypothesis that CAA will outperform the traditional or the existing complex performance assessment standard setting methods by comparing and analyzing the properties and assumptions of these methods. Then, we present the framework of the CAA as well as its properties. We finally exemplify CAA with a proposed standard-setting procedure and discuss its utility in real applications.

Rationale of Cognitive Analytical Approach

Validity of standard and cut score

Performance standards and cut score are defined as distinct but related concepts (Kane, 1994; Waltman, 1997). Performance standards refer to the minimally adequate level of KSAs that students must demonstrate for some purpose, while *cut score* is a point on a score scale that forms the boundaries between contiguous levels of student performances. The cut scores that differentiate examinees on performance levels define an ordinal scale that adds more interpretation to the existing information compared to raw scores or scale scores alone. The evaluative labels (i.e., PLL) defined by the cut scores suggest substantial differences between the performance levels. Examinees assigned with a PLL are assumed to have met the required KSAs described in the Performance Level Descriptor (PLD) corresponding to that PLL and should have demonstrated the evidence of that level of proficiency in the assessment. The appropriateness of the standards, cut scores, and the claims based on them need to be validated by the evidence shown in details in The Standards for Educational and Psychological Tests ([AERA, APA, & NCME], 1999). However, as noted by Kane (2001), like policy decisions, the standards are hard to validate, especially by comparing with external criteria, so are the consequences from the decisions of the standard setting.

We may never be able to set a “correct” cut score. Nevertheless, a clear set of performance standards makes it easier to state the PLDs and set the cut score. Kane (2001) has pointed out that procedural evidence was especially important in evaluating the appropriateness of performance standards and that the standards tend to be more convincing if they have been set in a reasonable way by knowledgeable people who know the process of standard setting and the purpose for which the standards are being set. To ensure the validity and defensibility of the standards, guidelines of standard setting were recommended to be used, which include the steps to select an appropriate standard-setting method, choose a panel, arrange the activities in the panel meeting, collect evidence of validity, and conduct technical analysis (Hambelton & Pitoniak, 2006; Cizek, 1996). The importance of building the link between the assessment and standard setting is stressed, for example, by choosing the standard-setting method based on the type of items or the computation of test scores, and connecting the standard-setting methods with KSAs being assessed (Hambleton & Pitoniak, 2006; Cizek, 1996).

Some researchers further took the stance of setting the standards before the tests were designed and administered (Bejar, et al., 2007; Bejar, 2008; Kane, 1994). Kane (1994) advocated specifying the performance standard and then developing the test according to the standards. Bejar et al. (2007) proposed creating the performance standard on an assessment framework that was consistent with the theories of diagnostic assessment design (Mislevy, et al., 2003). By this approach, it is more likely that the standards will cover the constructs of interest. They argued that this approach tended to lead to more valid and reliable standard setting results.

Cognitive diagnostic assessment design

The CAA standard setting approach requires a thoughtful integration of educational policy, learning theory and curricular considerations in the process of constructing a framework to guide the development of performance standards. Each of the steps requires judgment. By following this framework, the judgments and decisions can be based on logical, articulated models and credible evidence. The evidence-centered design (ECD) framework described by Mislevy and his colleagues (Mislevy & Haertel, 2006) is an overarching and systematic framework for diagnostic assessment design. ECD incorporates models of learning throughout the assessment process and simultaneously provides support for a systematic approach to standard setting and therefore we believe it to be more likely to lead to improved learning (Mislevy and Haertel, 2006).

ECD is aimed at providing an evidentiary argument for inference about what the examinees know, can do or have acquired from what we observe them say, do or make in a few assessment circumstances (Mislevy, et al., 2003; Mislevy & Haertel, 2006). The construct-centered approach advocated by Messick (1994) supports a chain of reasoning in ECD to construct a valid assessment and develop rational scoring rubrics. This approach consists of finding a representation of constructs related to instructions or societal values, behaviors or performances revealing those constructs, and the tasks or situations that elicit those behaviors. ECD applies to the processes of designing, implementing, and delivering an educational assessment. Its key concepts and entities, and knowledge representations and tools thread through the layers of domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. The layered framework of ECD affords intra-field investigations while simultaneously providing structures that facilitate communication across various kinds of expertise.

Domain analysis is intended to abstract substantive information of the concepts, terminology, and knowledge representation of the domain to be assessed. Many cognitive models provide a good starting point at this stage, for example, Bloom's taxonomy (Bloom, Engelhart, Fust, Hill, & Krathwohl, 1956) that differentiates learning into the hierarchical levels of knowledge, recall, application, analysis, synthesis, and evaluation, for another example, Anderson's ACT (Adaptive Components of Thought) theory that describes the phases of acquiring declarative knowledge and procedural knowledge respectively (Anderson, 1976;1983). *Domain modeling* adopts the terminologies and reasoning from Toulmin's diagram for the assessment argument, that is, providing an explanation of the claims about a student or his/her proficiency demonstrated by the tasks created from the design pattern. The *conceptual assessment framework* (CAF) consists of student models, evidence models and task models, where technical specifications are designated. A *Student model* expresses the KSAs that the assessment designer is intending to measure in a domain of tasks, a multidimensional student model for instance. A *task model* describes the environment that elicits the student behaviors to provide evidence. The *evidence model* connects the student model and the task model, namely, evaluating the information extracted from the work products through scoring and synthesizing the evaluation data to obtain the values on measurement variables through particular measurement models such as IRT.

Standard setting in the framework of ECD

The standard setting design proposed by Bejar et al. (2007) is characterized by taking account of performance standards in the early stage of ECD and developing the performance standards for several grades simultaneously. The articulation of performance standards at an early stage is important to inform the rest of the assessment development, in addition to serving as the basis for the cut scores that become the realization of those performance standards. A conceptual model is depicted in Figure 1. The essence of this approach is that the content standards (i.e., description of what students are expected to learn) and a competency model (i.e., mechanism of how students learn) inform the formulation of performance standards, which in turn can inform the test development process. Standard setting interacts with domain analysis and modeling. The content standards, the educational policy and the learning constructs are transformed into more concrete assessment elements, influencing evidence models and task models in CAF and consequently the test specifications and PLDs. In this way, standard setting is aligned with the framework set up by ECD. Cut-score setting is an iterative process that is subject to pragmatic and psychometric constraints, informed by the plausible theory-driven maximal discrimination region on the scale, tested by the field trials, and continuously adjusted by these earlier obtained data, as appropriate.

We agree that involving standard setting at an early stage of assessment design is an efficient approach to keep performance standards in line with the content standards as well as the cognitive framework. By this means, it is more likely to reduce the cognitive load for standard setting in complex performance tasks and ensure the validity of the test development and performance standards. Therefore, we use this approach to guide our CAA standard setting.

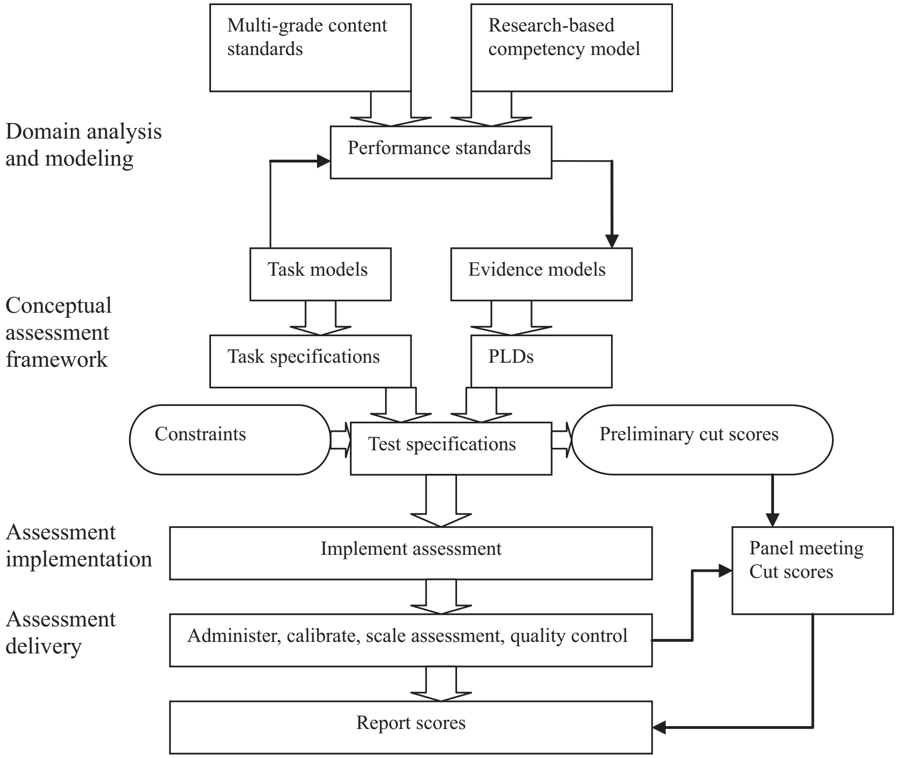
Multi-grade content standards
Research-based competency model
Domain analysis and modeling
Performance standards
Task models
Evidence models
Conceptual assessment framework
Task specifications
PLDs
Constraints
Test specifications
Preliminary cut scores
Assessment implementation
Implement assessment
Panel meeting Cut scores
Assessment delivery
Administer, calibrate, scale assessment, quality control
Report scores

Figure 1:
Flowchart of standard-setting in the framework of evidence-centered design
(adapted from Bejar, et al., 2007)

CAA compared with other standard setting methods

CAA is meant to be implemented simultaneously with content specification and test development. It is designed to capture the non-compensatory, multiple dimensions for performance assessment. There might be no resolution to the argument about whether CAA is adequate or appropriate for a test, because it is hard to quantify the personal and societal costs and benefits associated with any particular performance standard (Kane, 1994). It might be virtually impossible to validate a claim that any performance standard is correct, but by validation we can justify that one standard-setting method is better than any of the others (Cizek, 2001). Five types of evidence are usually considered to evaluate the validity: explicitness, practicability, implementation of procedures, panelist evaluations and documentation (Hambleton & Pitoniak, 2006).

Traditional standard setting methods

Traditional standard-setting methods are usually classified into a dichotomy: test-centered methods versus examinee-centered methods (Cizek, 2001). Test-centered standard-setting methods require panelists to make judgments on the expected levels of performance by borderline examinees on each assessment task (e.g., Angoff, modified Angoff, bookmark); while examinee-centered standard-setting methods require panelists, who know the students, to place the students into performance categories, without any knowledge about their actual performance on the test (e.g. contrasting groups). This dichotomy can be applied as follows to three very popular approaches to standard setting:

1. Modified Angoff: the judges estimate the probability that a minimally proficient or minimally advanced student will get the item correct. One alternative approach is to rate the item as more likely to be answered correctly or incorrectly by a student who is minimally proficient or is minimally advanced. This procedure assumes that ordering items by the probability of getting an item correct (difficulty level) is also ordering the level of KSA. The sum of the numbers or probabilities across all the items in the test is the cut-point as determined by that standard setter for that test. Averaging across standard setters provides the recommended cut-score for each of the three levels of student performance. Cut scores are set iteratively. In each round, the standard setters are usually informed about the impact data, that is, how the cut scores they have recommended are going to affect the classification of the population of students who have taken or will take the test.
2. Bookmark: a number of items are examined and are organized from easiest to most difficult by the p value in classical testing theory or item difficulty parameter b value in IRT. The task for the standard setting panelists is to place a "bookmark" between the hardest items that a basic student would get right and the easiest item that the basic student would not get right. Again, this approach asks the standard setters to use the PLDs to determine the placement of the cutoff and their work is informed by discussion with other members of the standard setting team. The difficulty of each item is central to this procedure and organizes the items by difficulty in what is called an "ordered item booklet."
3. Traditional contrasting groups method: the judges, who are familiar with a group of examinees, are asked to use the PLDs to identify examinees who are clearly above a particular performance standard and those who are apparently below that performance standard based on their knowledge about the examinees' overall performances or proficiencies. Then the test score distributions of these two groups are plotted and the cut score is placed at the point where the two distributions intersect (Cizek, 2001). Ordering by test scores implies, again, a reliance on the difficulty of the test items that aggregate to that total score to define the KSA of the construct(s) for which the standard is being determined.

It is noteworthy that the traditional standard setting methods have in common a single scale along which either the item difficulties or the levels of ability are rank ordered. The item difficulty can be the p-value or the IRT difficulty parameter, response probability or an average item score for constructed response items. This scale is analogous to the item difficulty/ability scale in IRT. For an examinee-centered standard setting such as the contrasting group method, the students are ordered by and within PLDs along a single continuum of the skill. The assumption is that the total test score is a monotonic function of the latent ability. In other words, students with higher value on the latent ability will score higher on their performance based on the total test. While for test-centered approaches, such as the bookmark method, items are placed along a single continuum of difficulty and a marker is placed to differentiate students who are able to answer items difficult enough to be considered proficient, yet not able to answer items so difficult as to be considered expert. If the items are assumed to be monotonic, they make sense lined up against the continuum implied by and within the PLDs.

This single scale embedded in the traditional standard setting provides a means of communication to panelists. The panelists must consider the ability of the students along a continuum that is adequately captured by difficulty or some essential variant of difficulty. It is implied that the placement of a student in a PLL should depend upon the difficulty of the items that he or she can answer correctly or with higher probability. It suggests that difficulty is a proxy for or monotonic to the ordering of the PLLs from the lowest level (e.g., being able to answer easiest items) to the highest level (e.g., being able to answer hardest items). This ignores the fact that there could be and almost certainly are multiple scales underlying a complex performance task. Even a composite scale cannot precisely capture several attributes at one time, unless they at least mirror each other monotonically. Cognitively simple knowledge level items can be very difficult for a variety of reasons and in fact might be much harder than more complex reasoning items.

Finally, there is considerable evidence that difficulty is not the same as cognitive complexity, and it is cognitive complexity that is at least the conceptual focus of standard setting. In other words, schools are not usually interested in whether students can answer “hard” knowledge items rather than analysis and synthesis items. It is the difference between students who operate cognitively at the knowledge level versus those that operate at more advanced cognitive levels that is of interest. Several papers have shown that assessment items that may be ordered in difficulty, do not necessarily order the same way by their cognitive complexity. Papers by Arend, Colom, Botella, Contreraa, Rubio, Snatacreu (2003), Spilsbury, Stankov and Roberts (1990), Stankov (2000) and Stankov and Raykov (1995) are examples of work in that area.

Standard setting methods for complex performance assessment

Over the last 10 years, many assessment programs have added constructed-response (CR) items, with a hope to deliver a test that is closer to real learning situations. CR questions require the examinees to produce the response in their own words. CR questions vary in cognitive and format complexity to a larger extent than multiple-choice questions. For instance, complex CR questions could require examinees to integrate knowledge and apply to a real-life situation, or provide a rationale to justify their responses.

One example in which a CR item is scored multidimensionally is the trait scoring for some writing assessments (e.g., writing test in Arizona Instruments to Measure Standards (AIMS)). A set of rubrics is created for latent traits (i.e., idea/content, organization, voice, word choice, sentence fluency, conventions). The answer is scored by rater's judgments regarding the performance on each trait. However, when it comes to the standard setting, a composite scale is created by averaging the trait scores to allocate an overall cut score. This provides no classification information on each of the traits for diagnostic purposes.

The new test format presents the need for appropriate standard setting methods to accommodate such complexity. Presented in this section are methods that could deal with tests containing constructed-response items. These methods involve either review of candidate work or review of score profiles. For the former type of standard setting, the product work could be viewed either item by item (Loomis & Bourque, 2001), or section by section (Plake & Hambleton, 2001; Plake, Hambleton, & Jaeger, 1997), or holistically (Jaeger & Mills, 2001), depending on the properties of the test, such as the type of items, the total number of CR questions, the complexity of the questions whatever type they are, or the actions required by the examinees (Cizek, 2001).

1. Item by item approaches: for each question, panelists are asked to select from a set of examinee performances the work that best represents the performance of minimally competent candidates. In some cases, the actual scores assigned to the papers are revealed to the panelists. Then the panelists make an estimate of the performance of the minimally competent candidate on each question. These standards are then aggregated to obtain the overall performance standard for the full test. However, since this approach takes place after the test is administered, it may be difficult for the panelists to adjust their performance standards from round to round. On the one hand, it may not be easily interpretable to the panelists how one score may represent borderline performance at a given performance standard, and another score represents borderline performance at another level. On the other hand, there may be a lack of papers at a given score point. Therefore, it may take a long time to prepare work representative at different levels.

2. Holistic approaches: Like the *Body of Work* (Kingston, et. al., 2001) they require panelists to view the samples of examinee performance holistically. Panelists are provided with more examinee papers representing a more focused score range around a cut point. The values in this range suggest where the minimum performance standard would be likely to fall. The score point where panelists seem indifferent to pass-fail decisions is chosen as the passing score. This process is repeated for each performance standard of interest. The limitation of this type of methods is that there is a maximal booklet length beyond which panelists cannot make valid and reliable judgments about the materials (Hambleton, Jaeger, et al., 2000). The researchers had observed that when panelists were presented with the complete work of examinees, they tended to skip over some of that work and key in on selected questions or the first part of the students' work.

3. Hybrid approaches: *Analytic judgment method* for instance, the panelists' ratings are based on components of the test, rather than on the entire test. Breaking up the test book-

let into smaller collections of test items was done to reduce the cognitive complexity of the rating task by reducing it to judging more modest sets of items. Panelists sort candidate papers into ordered performance categories. The ratings can be transformed into performance standards by using a boundary method (i.e., averaging the scores of papers assigned to the high end of one performance category and the low end of the next higher performance); the performance standards established for each set of test items are then summed in order to obtain performance standards for the total test. However, this set of approaches does not set standards for multiple dimensions in particular. Again, the procedure depends upon the reasonableness of adding scores and that depends upon their being at least monotonically related.

We may notice that no matter how the work is reviewed item-by-item or holistically, scores assigned to the performance of borderline candidates are ultimately aggregated across the test and result in a set of standards to evaluate the overall performance. Proficiency is measured on a composite scale that is directly related to number correct or some weighted average or sum of sub-scores. In contrast, methods that involve review of the score profile address the standard setting for the complex exercises that are scored multidimensionally and focus on the cognitive structure that underlies the test. That focus is retained as long as the process does not involve adding the multidimensional scores together to form a single composite.

In the *Judgmental policy-capturing (JPC) method,* the panelists' task is to review hypothetical score profiles and rate a large number of vectors of scores, and the standards are inferred from a statistical analysis of their ratings. In one of the variations implemented for the National Board for Professional Teaching Standards Certifications (Jaeger, 1995), each exercise and the entire assessment were scored multidimensionally. The panelists were provided with information about their own ratings of profiles and the ratings of the entire panel. This approach was claimed to be feasible and reliable (Jaeger, 1995b), but Hambleton (1998) also noted that it is challenging to find statistical models that fit the panelists' ratings and explain the overall process to panelists for deriving a performance standard. *Dominant profile method (DPM)* is another approach where the panelists, who are fully aware of the questions and the meaning of the scores, derive decision rules that capture the score levels across the profile components. With a large number of possible score profiles, it is hard to reconcile panelists' views into a consensus.

Methods and procedures

The following is an illustrative example that we have created around the standard setting flowchart proposed by Bejar et al. (2007). Using CAA, we demonstrate how one might establish the performance standards, task models, and evidence models before defining the task specifications and PLDs, which in turn precede formulating the test specifications and item development. Setting cut scores is now an iterative process along with test construction and does not depend upon test performance or upon difficulty level or its aggregation. When agreement is attained on task models and constraints, the blueprint of the test specification can be finalized, and the preliminary cut scores corresponding to the performance standards are determined. Preliminary cut scores can be evaluated after the assessment implementation and adjusted in light of the data available, if it is desired to do so, but that is not necessary. Impact data are often of great interest to policy considerations, but are not of much interest at a conceptual level. The standards could be specified across the grades by systematically basing new learning on the preceding acquired skills, but here we focus on the CAA procedure for one grade to illustrate its process and application.

Purpose of the assessment

The current standard setting is assumed to take place in a large-scale performance assessment that is intended to diagnose the Knowledge, Skills and Abilities (KSA) in mathematics for regular students in 6th grade. Students are required to draw on a broad body of mathematical knowledge and apply a variety of mathematical skills and strategies. In order to function as a citizen and a worker in the contemporary society, a person should have the ability to explore, to conjecture, to reason logically, to communicate in mathematics effectively, and to apply a wide repertoire of methods to solve problems.

Domain analysis

Through the analysis, we have the following list of content-related standards for 6th grade: (1) numbers and operations, (2) data analysis, probability and discrete analysis, (3) patterns, algebra and functions, (4) geometry and measurement, (5) structure and logic. Other abilities we call structural KSAs are (1) communication, (2) problem-solving, (3) reasoning proof, (4) connections, and (5) representation that are embedded throughout the teaching and learning of all mathematical content.

Content standards

Learning objectives represent the expectations in regard to each content area. The skills necessary to meet those expectations are identified. We take the Measurement component in Geometry and Measurement for example (Table 2). Key skills required for each objective are listed, some of which come from the previous learning objectives. For example, to estimate the measure of objects using a scale drawing or map, the required KSAs are, in brief, the content-related KSA of fractions, and the structural KSAs of problem-solving, reasoning, representation and communication.

Proficiency level descriptors (PLDs)

PLDs identify the evidence that is determinant to the proficiency levels. The evidence can evolve from an abstract expectation to a more concrete form of descriptions for “advanced”, “proficient” and “basic” levels. As is shown in Table 3, the PLDs are labeled first in an abstract form, and then in a concrete form as in Table 4. The set of all PLDs corresponding to the learning objectives are derived, but not all the learning objectives are covered in Table 4. Three sets of PLDs are shown for three different proficiency levels, but notice that these are for a specific skill/concept and there may be many such in a test to which CAA is applied.

Test specifications

As the PLDs are elaborated it is necessary to create tasks that can be expected to elicit evidence linked to the PLDs. Tasks could take forms ranging from multiple choice items to open-ended questions. The task model can be built upon the structural variables and content-related variables defined in the content standards. Each of the task models can be represented in a variety of ways. Given a specific instance of a task model we can describe the structural attributes, including the PLDs, that the task was designed to elicit. One of the conveniences brought by CAA is to designate individual items to discriminate between the adjacent proficiency levels before data collection and analysis. Another advantage is to specify the KSAs involved in the design of a particular item. At the early stage of developing CAA, we assume that each item is rated with a score vector on the designated KSAs assessed. At this stage, we may focus on the test tasks where KSAs are non-compensatory (the score on one KSA is independent of the score on another one) in accomplishing a correct answer to the item. An example of test specification is presented in Table 1, a possible structure of test specification. The analysis inherent in CAA might even suggest that additional items need to be written to permit more accurate measurement associated with specific score vectors.

Table 1:
Table of test specifications

	KSA1	KSA2	KSA3	KSA4	……	KSAm
Item 1	A/P		P/B			
Item 2		A/P		P/B		A/P
Item 3		A/P	A/P			
Item 4	P/B			A/P		
Item 5	A/P					
Item 6						P/B
⋮						
Item n		P/B	A/P			

A=Advanced, P=Proficiency, B=Basic

Table 2:
Learning objectives and the key KSAs. ([4]Arizona mathematics standard articulated by grade level, grade 6, 2008)

Strand 4: Geometry and Measurement Geometry involves the development of students' reasoning, higher-order thinking, and justification skills culminating in work with proofs. Geometric modeling and spatial reasoning offer ways to interpret and describe physical environments and can be important tools in problem solving. Students use geometric methods, properties and relationships, transformations, and coordinate geometry as a means to recognize, draw, describe, connect, analyze, and measure shapes and representations in the physical world. Measurement is the assignment of a numerical value to an attribute of an object, such as the length of a pencil. At more sophisticated levels, measurement involves assigning a number to a characteristic of a situation, as is done by the consumer price index. A major emphasis in this strand is becoming familiar with the units and processes that are used in measuring attributes.	
Measurement Understand and apply appropriate units of measure, measurement techniques, and formulas to determine measurements. In Grade 6, students build upon their prior knowledge of measurement to determine the appropriate unit of measure, tool, and necessary precision to solve problems. They convert within systems of measurement to solve problems. They use scale drawings to estimate the measure of an object. Students also apply formulas for area and perimeter to solve problems and explore the relationship between volume and area.	
Performance Objectives	***Key KSAs***
Students are expected to:	
PO 1. Determine the appropriate unit of measure for a given context and the appropriate tool to measure to the needed precision (including length, capacity, angles, time, and mass). Connections: M06-S1C3-02, SC06-S1C2-04	*(M06-S5C2-01) Analyze a problem situation to determine the question(s) to be answered. (M06-S1C3-02) Multiply and divide fractions. (SC06-S1C2-04) Perform measurements using appropriate scientific tools (e.g., balances, microscopes, probes, micrometers).
PO 2. Solve problems involving conversion within the U.S. Customary and within the metric system. Connections: M06-S1C1-03, M06-S1C3-02	*(M06-S5C2-03) Apply a previously used problem-solving strategy in a new context. (M06-S1C1-03) Demonstrate an understanding of fractions as rates, division of whole numbers, parts of a whole, parts of a set, and locations on a real number line. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results.

[4] http://www.ade.state.az.us/standards/math/Articulated08/Gradeleveldocs/MathGrade6.pdf

PO 3. Estimate the measure of objects using a scale drawing or map. Connections: M06-S1C1-03, M06-S1C3-02, SS06-S4C1-03	*(M06-S5C2-03) Analyze and compare mathematical strategies for efficient problem solving; select and use one or more strategies to solve a problem. (M06-S1C1-03) Demonstrate an understanding of fractions as rates, division of whole numbers, parts of a whole, parts of a set, and locations on a real number line. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (SS06-S4C1-03) Interpret maps, charts, and geographic databases using geographic information
PO 4. Solve problems involving the area of simple polygons using formulas for rectangles and triangles. Connections: M06-S1C3-02, M06-S3C3-04, M06-S5C1-02	*(M06-S5C2-02) Identify relevant, missing, and extraneous information related to the solution to a problem. *(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (M06-S3C3-04) Evaluate an expression involving the four basic operations by substituting given fractions and decimals for the variable. (M06-S5C1-02) Create and justify an algorithm to determine the area of a given compound figure using parallelograms and triangles.
PO 5. Solve problems involving area and perimeter of regular and irregular polygons. Connections: M06-S1C3-02, M06-S3C3-04, M06-S5C1-02	*(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context. (M06-S1C3-02) Make estimates appropriate to a given situation and verify the reasonableness of the results. (M06-S3C3-04) Evaluate an expression involving the four basic operations by substituting given fractions and decimals for the variable. (M06-S5C1-02) Create and justify an algorithm to determine the area of a given compound figure using parallelograms and triangles.
PO 6. Describe the relationship between the volume of a figure and the area of its base.	*(M06-S5C2-04) Apply a previously used problem-solving strategy in a new context.

Table 3:
Performance level descriptors on the general KSAs

Performance Level	**Descriptor**
Advanced	The student exceeds the expectations for demonstrating an independent and accurate understanding of the specified math skills/concepts. The student demonstrates the ability to apply the skills/concepts to an authentic task and/or environment with analysis and reflection by: – solving a real world problem (e.g., determining what fraction of a dozen eggs are needed to bake a cake if 3 are needed) – applying math skill/concept in the natural environment (e.g., store, home, technical education class, science class, home economics, etc.) to solve a problem – communicating an in-depth explanation that analyzes or reflects on the problem (e.g., demonstrate how left over pieces of one pizza can be combined with pieces of another pizza to create a whole pizza and explain how that works)
Proficient	The student demonstrates an independent and accurate understanding of the specified math skills/concepts. Occasional inaccuracies, which do not interfere with conceptual understanding, may be present. The student demonstrates the ability to apply the skills/concepts to an authentic task and/or environment by: – solving a real world problem (e.g., determining what fraction of a dozen eggs are needed to bake a cake if 3 are needed; determine the perimeter of a table to determine the amount of ribbon needed to decorate the sides; reproduce two dimensional shapes to complete an art project; construct a bar graph showing class election results; etc.) – applying math skill/concept in the natural environment (e.g., store, home, technical education class, science class, home economics, etc.) to solve a problem. – using relevant details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – using math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.) – using a model or explanation to demonstrate a concept or solve a problem (e.g., create a chart showing fractional parts; draw a floor plan of a clubhouse and provide area; categorize shapes according to elements; create a bar graph and answer questions; etc.)
basic	The student demonstrates basic understanding of the specified math skills/concepts. Inaccuracies may interfere with or limit the conceptual understanding. The student demonstrates some understanding without applying the skills/concepts to an authentic task and/or environment by:

	– solving a problem (e.g., identify fractions on worksheet, figure area problems; match element to 2 D shape; complete numerical pattern; etc.) – using relevant details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – using math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.) **or by:** – using a model or explanation to demonstrate a concept or solve a problem (e.g., create a chart showing fractional parts; draw a floor plan of a clubhouse and provide area; categorize shapes according to elements; create a bar graph and answer questions; etc.)
Below basic	The student demonstrates little or no understanding of the math skills/concepts. Inaccuracies interfere with the conceptual understanding. The student demonstrates this by: – inaccurate use of details (e.g., uses measurements, elements of 2 D shapes, data, numbers, etc.) – inaccurate or no use of math vocabulary (e.g., fractions, whole, area, perimeter, rectangle, square, data, graph, pattern, etc.)

Table 4:
[5]Arizona mathematics standard performance level descriptors on specific learning objectives (grade 6)

Students at the “Advanced” level generally know the skills required at the “Proficient” and “Basic” levels and are able to:	Students at the “Proficient” level generally know the skills required at the “Basic” level and are able to:	Students at the “Basic” level generally know and are able to:
• Use prime factorization to determine greatest common factor and least common multiple. • Express the inverse relationships between exponents and roots for perfect squares and cubes. • Apply and interpret the concepts of addition and	• Convert between fractions, decimals, percents, and ratios. • Express a whole number as the product of its prime factors. • Demonstrate an understanding of fractions as rates or as division of whole numbers. • Compare and order integers, positive fractions, positive decimals, and positive percents.	• Express that a number’s distance from zero on the number line is its absolute value. • Apply properties to solve numerical problems. • Make estimates appropriate to a given situation and verify

[5] http://www.ade.state.az.us/standards/AIMS/PerformanceStandards/6thMathPLD.pdf

subtraction with integers using models. • Provide a mathematical argument to explain operations with two or more fractions or decimals. • Build and explore tree diagrams where items repeat. • Investigate and solve problems using Hamilton paths and circuits. • Create and solve two-step equations with fractions and decimals. • Solve problems involving supplementary, complementary, and vertical angles. • Solve problems involving area and perimeter of regular and irregular polygons. • Describe the relationship between the volume of a figure and the area of its base. • Create, analyze, and justify algorithms for multiplication and division of fractions and decimals and area of compound figures. • Make and test conjectures based on information collected from explorations and experiments. • Solve simple logic problems and justify solution methods and reasoning.	• Multiply and divide decimals or fractions. • Simplify numerical expressions using the order of operations. • Use benchmarks as meaningful points of comparison for rational numbers. • Interpret, describe, and analyze displays of data. • Determine theoretical probability and apply it to predicting experimental outcomes. • Analyze numerical patterns using all four operations. • Describe the relationship between two quantities in a function. • Use an algebraic expression to represent a quantity. • Evaluate an expression by substituting given fractions and decimals for the variable. • Solve problems involving the relationship among the circumference, diameter, and radius of a circle. • Identify the missing coordinate of a polygon on the coordinate plane. • Solve problems involving conversion within the U.S. Customary and within the metric system. • Solve problems involving the area of simple polygons using formulas for rectangles and triangles. • Evaluate situations and select strategies to find and apply solutions to problems. • Compare sets of data by analyzing trends. • Explore counting problems using Venn diagrams with three attributes.	the reasonableness of the results. • Identify a simple translation or reflection of a 2-dimensional figure on a coordinate plane. • Graph ordered pairs in any quadrant of the coordinate plane. • Determine the appropriate unit of measure for a given context. • Estimate the measure of objects using a scale drawing or map.

Panelists' selection

We assume that the panelists consist of teachers, non-teacher educators, test developers, and the general public. Panelists are usually recruited statewide through a stratified sampling, and we will assume that has occurred in this hypothetical example. In many standard setting applications, sampling would try to have no less than 30% of the panelists from ethnic minority groups and no less than 25% of them being males. In this case, we will retain the usual three rounds of panelists' meetings to set cut scores, once before the tests are developed and once after the tests are written and administered. Notice, that unlike the usual standard setting, these rounds are separated by long time periods of intense test related work. The cut scores are finalized as the third and last round of discussions. For illustrative purposes, let's assume that 18 panelists are distributed to three tables with 6 at each table and stratification is utilized to maintain balance along various dimensions of interest such as race, gender, geographic region and SES level of the typical student at the participant's school.

Orientation and discussions

Round 1. Panelists receive an overview of CAA method in a 60-minute presentation. The presentation describes the purpose of the diagnostic assessment, the basic concepts and framework of ECD, and interpretations of PLDs. The role of standard setting prior to the test development is explained and its value and interpretation is made clear.

KSA Review. Panelists are presented with a learning objective table such as Table 2, showing learning objectives in each content domain area and the KSAs necessary to meet a learning objective. When the KSA review is complete, panelists should have a detailed, structured understanding of the assessment and expected student achievement.

PLD Review. Panelists also review the PLD tables in both abstract (Table 3) and concrete (Table 4) forms. The abstract PLDs describe the expectations on general KSAs (e.g., analyzing, application, problem-solving and communication), which are less related to the specific content. In contrast, the concrete form provides PLDs on a sample of learning objectives.

Test Specifications Review. Panelists are also instructed to study the test-specification table as shown in Table 1, where the items are represented as capable of discriminating between adjacent performance levels on KSAs based on the attributes of the learning objectives. Panelists are asked to think of a task, preferably in the form of an item that exemplifies the content knowledge, skill or ability given in the specifications. Panelists are also asked to share items with the whole group for discussions.

Preliminary cut-score setting. With clear test blueprints such as summarized in Table 1, the next step is to obtain preliminary cut scores. At this stage in the development process, prior information on the PLDs has been accumulated and, moreover, the PLDs can be associated with the learning expectations linked to the performance levels. Based on the characteristics on the KSA continuum, there are items that are more likely to discriminate between “advanced” and “proficient”, and items that are likely to discriminate between “proficient” and “basic”. It is feasible, and desirable, to associate performance levels with possible performance on a test, even though the test has not been fully implemented or administered.

Panelists spend the next five hours of meeting time identifying the knowledge, skills, and abilities, and the learning objectives students must have to qualify for “advanced” or “proficient”. They also read the test specifications as presented in Table 1. For each KSA, panelists can make their decisions on the cut scores by aggregating the ratings on the same KSAs across items (i.e., calculating the number of “A” or “P” or “B) and fill in a cut-score table (Table 5). The specification table may be revised if more or less information on a particular KSA is needed. For example, the cut-score of “proficient” and “basic” is all “Ps” on problem-solving, and the cut-score of “advanced” is at least four out of five “As” and one “P” on the same KSA.

Test Development. Test developers generate tasks that best discriminate the levels designated in the table of specifications and written items. Tests are administered and scored on KSAs (the score points differentiate “advanced” and “proficient”, “proficient” and “basic”). For example, for the learning objective “to determine the appropriate unit of measure for a given context and the appropriate tool to measure to the needed precision (including length, capacity, angles, time, and mass)”, panelists, in their first round of discussion, decide that this would involve the content-related KSAs of fractions and using measurement tools, and the structural KSAs of problem-solving and reasoning. The KSAs on fractions and measurement tools are likely to discriminate the proficient students from the advanced ones, while on problem-solving and reasoning, these items are expected to differentiate “proficient” and “basic” students. Based on these task features, an item could be constructed as follows: In your science class, you want to measure leaf width and plant heights to determine the effects of different kinds of fertilizers. What tools and units of measure would you use to make the measurements? To what degree of precision should you measure? Explain and justify your choices.

Round 2. Panelists are convened again after the test design implications from round one are implemented and they have a brief review on the KSAs that each item measures. They are presented with sample papers with a wide range of proficiency levels. The panels, again, keeping in mind the performance level descriptors on each KSA and using a table like Table 1 to rate the performance as “A”, or “P” or “B” for each KSA, decide the minimum number of “As”, “Ps” and “Bs” for each proficiency level of a KSA (Table 5).

Table 5:
Cut-score table

	KSA1	KSA2	KSA3	KSA4	……	KSAm
Basic/proficient	4Ps&1A	6Ps	5Ps & 1B	7Ps		5Ps&2As
Proficient/Advanced	5As	5As&1P	6As	5As&2Ps		6As&1P

Round 3. Cut-score results from the Round 1 and Round 2 are provided for comparison purposes. Panelists are shown the numerical values of the Round 1 and Round 2 medians. Panelists could see the change in the median from Round 1 to Round 2, and give cut score recommendations. This is an iterative process. Discussions take place to explore and try to resolve salient differences of opinion within each group. Panelists will be provided with results from other groups and discussions will continue until a consensus is (hopefully) achieved for the whole group.

The procedure illustrated above is one of the possible procedures of CAA. Other variations could be a bookmark like procedure that orders sets of items along the specific scales of KSA and places a bookmark at the borderline that divides the proficiency levels for each KSA, or an Angoff procedure that requires judgments on the probabilities of correct answer for the minimally proficient candidates, again on the relevant items measuring a specific scale. Notice that the essential multidimensionality of the test is maintained in the standard setting. The procedure illustrated above in detail represents a hybrid approach that integrates both a test-centered component in Round 1 and an examinee-centered component in Round 2. This hybrid approach enables the performance standards to be determined in what we argue as a more sensible way. Other variations on the essential ideas of CAA can be implemented, as the client (state) might choose.

Discussions and conclusions

In CAA, the performance standards are established simultaneously with domain modeling and test specifications; the standards and cut scores are evaluated iteratively along with the test design and development phases. CAA has the benefits of ensuring the validity of the performance standards, reducing the cognitive load of standard setting, including the complexity of the tasks, and facilitating the vertical articulation of KSAs. In this paper, we elucidate the theoretical and practical rationale of CAA and demonstrate its procedures and results with an illustrative example that we have created to show how this process might unfold.

CAA that is specifically tailored for cognitive diagnostic assessment is a thoughtful integration of educational policy, learning theories and curricular considerations in the process of constructing a framework to guide the development of performance standards. At the first stage, the learning objectives are translated into proficiency models and then linked to PLDs. The standards are set in regard to each learning objective while the test specifications are also determined. Once the tests are created and implemented, judgment is required again to reevaluate the performance standards and transform them into a set of cut scores. One of the major advantages of this approach is that with the guidance of ECD, the cognitive structure is maintained to be consistent and coherent across the stages from the domain modeling to score reporting. By this means, we would have more convincing evidence for the construct relevant validity since the test is designed to adhere to this structure.

CAA is innovative and appropriate for the cognitive diagnostic assessment compared with the existing standard setting methods. The traditional standard setting methods assume a unidimensional scale along which the abilities or the item difficulty values are rank ordered. This simplified cognitive pattern facilitates the communication with the standard setters, but it is an incorrect and misleading assumption with respect to the latent structure for complex performance tasks tapping into multiple skills. The contemporary standard setting methods for complex performance assessments which fall in the categories of analytical or holistic methods treat each item as a distinctive instrument measuring a sub-domain of skills, but items are still assumed to be unidimensional and the standard setting procedures result in an overall cut score on the composite scale that is a fiction. The current standard setting methods that involve the creation and review of score profiles tend to result in a large number of score patterns, which make it cognitively challenging to reduce to a smaller number of performance standards that are conceptually sensible.

In contrast, CAA considers a pattern of constructs to be assessed at the very beginning, and designates the constructs to determine the test specifications. CAA becomes an integral part of the planning and design process. In other words, the dimensions and their standards are designed into the test at the beginning. The performance levels that each item is intended to discriminate are specified as part of the development process. This approach recognizes the multidimensional latent structure of CR items and MC items and facilitates setting cut scores on several constructs at a time. The participation of test developers helps to ensure the consistency of assessment design and standard setting. The standards are set in a way consistent with how the learning objectives are labeled and items are scored. That is, the panelists are able to express their standards in terms of the number of “As”, “Ps” or “Bs”, which is explicit and determined prior to any test administration and data collection. In addition, CAA provides a systematic approach to develop the standards for different grades, and thus has the potential for setting standards across grades. We have not explicitly addressed this application, but creating panels from different subject matter areas and especially different grades can be used to create vertically moderated standards (Lissitz and Huynh, 2003).

We take account of the assessments with CR items in this study. Further research could investigate complex performance assessments that involve both multiple choice and CR items, make a distinction between the different test formats and update the standard setting methods accordingly. We could also examine the utility of other variations of CAA that use bookmark or modified Angoff procedures adapted for this purpose.

Some researchers (Roussos, et al., 2007) proposed model-driven classifications using probabilistic diagnostic models to estimate the cut scores to classify the students at different levels. On the one hand, this is an objective approach to obtain the classifications from the data and model. On the other hand, some of the parameters in the diagnostic models are specified based on the cognitive theory, such as those in the Q matrix that connect the latent attributes and the items, and many other assumptions are imposed to make the estimation possible. In addition, model identification will be an issue especially for a small-scale performance assessment where the examinee pool is not big enough to ensure all parameters can be accurately estimated. Importantly, the probabilistic diagnostic models are grounded in probability theory and applications of Bayesian statistics and might not be accessible or interpretable for most of the audiences that receive the score reports or the classification results. Finally, such models are usually implemented after the test data are obtained and our approach is designed to be a part of the test construction process. CAA can be regarded as complementary to the probabilistic diagnostic approach. They are both based on a certain kind of cognitive diagnostic framework, but through different classification procedures. However, it would be interesting to compare the results of the standard setting by human judgment with the model-driven classifications.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, J. R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Arend, I., Colom, R., Botella, J., Contreras, M.J., Rubio, V., & Santacreu, J. (2003). Quantifying cognitive complexity: evidence from a reasoning task. *Personality and Individual Differences. 35*(3), 659-669

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. R. Lissitz (Ed.), *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

Bejar, I. I. (2008). Standard setting: What is it? Why is it important? R&D Connection. 7. www.ets.org

Bloom, B. S. (ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: Handbook I: Cognitive Domain*. New York: David McKay

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice, 15*(1), 12-21.

Cizek, G.J. (2001).Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of Cognitive Diagnostic Assessment and a Summary of Psychometric Models. *Handbook of Statistics*, 26, 1-52.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.

Goodman, D. P. & Hambleton, R. K. (2004). Student Test Score Reports and Interpretive Guides: Review of Current Practices and Suggestions for Future Research. *Applied Measurement in Education, 17*(2), 145-220.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355-366.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.433-470). Westport, CT: Praeger.

Jaeger, R. M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

Jaeger, R. M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, 14*(4), 16-20.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational* Research, 64(3), 425-461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170.

Kingston, N. M., Kahl, S. R., Sweeney, & Bay, L. (2001). Setting Performance Standards Using the Body of Work Method. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp. 219-248). Mahwah, NJ: Lawrence Earlbaum Associates.

Lissitz, R. W., & Huynh, H. (2003). *Vertical Equating for State Assessments: Issues and Solutions in Determination of Adequate Yearly Progress and School Accountability*. Practical Assessment Research and Evaluation.

Lissitz, R. W., & Li, F. (2010). *Standard setting in complex performance assessments: An approach aligned with cognitive diagnostic models*. National Council on Measurement in Education, Denver.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed*.), Standard setting: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 1994(23), 13-23

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek. (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Lawrence Erlbaum.

Plake, B. S, Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement, 57*, 400-411.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 275-318). New York, NY: Cambridge University Press.

Spilsbury, G., Stankov, L., & Roberts, R. D. (1990). The effect of a test's difficulty on its correlation with intelligence. *Personality and Individual Differences, 11*(10), 1069-1077.

Stankov, L. (2000). Complexity, Metacognition, and Fluid Intelligence. *Intelligence, 28*(2), 121-143.

Stankov, L., & Raykov, T. (1995). Modeling complexity and difficulty in measures of fluid intelligence. *Structural Equation Modeling, 2*(4), 335-366.

Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement, 34,* 101-121.