STAGE 7

Standard Setting

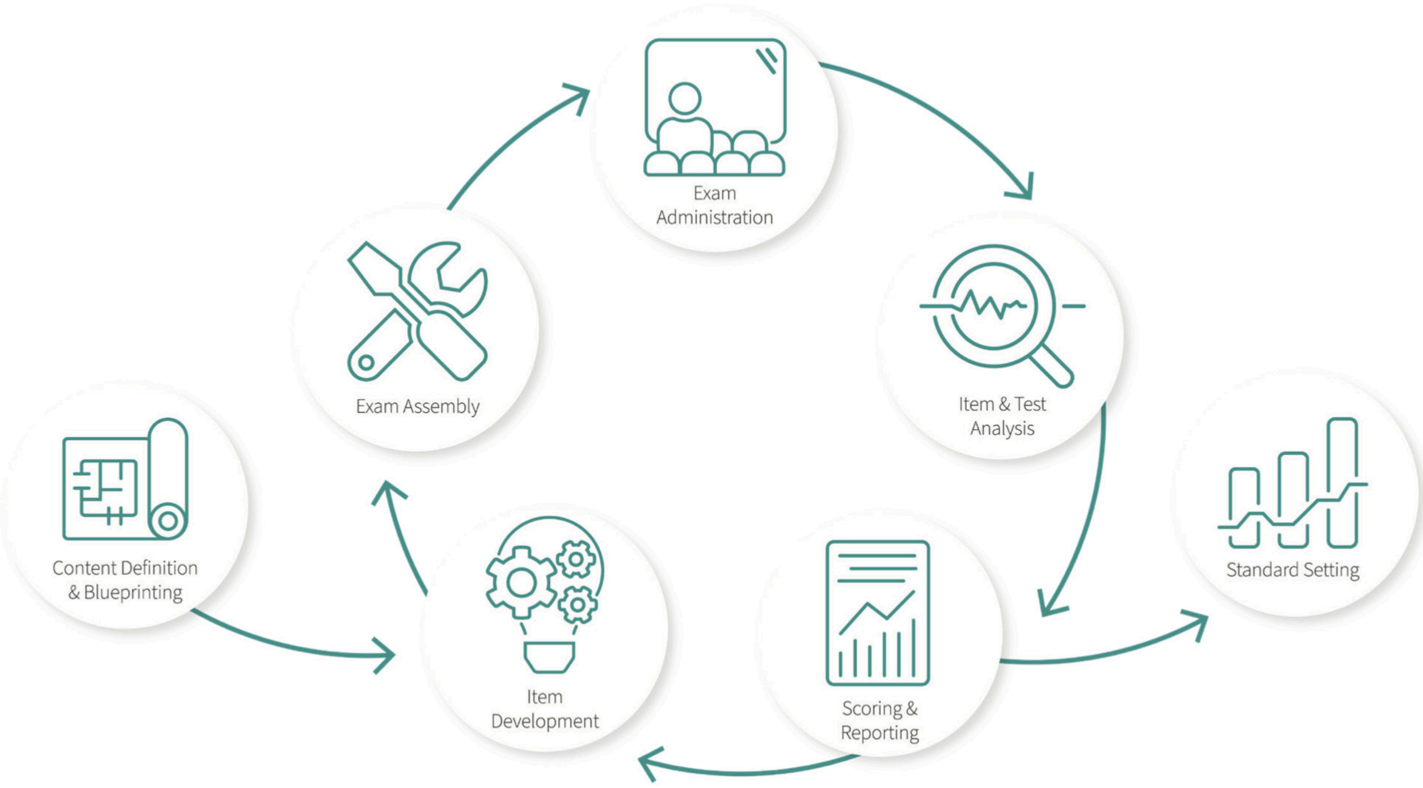
MEAZURE LEARNING

Meazure Learning Whitepaper

Table of Contents

Introduction

The Assessment Life Cycle is a way of organizing the processes involved in creating valid assessments into a series of easy-to-understand, logical stages.



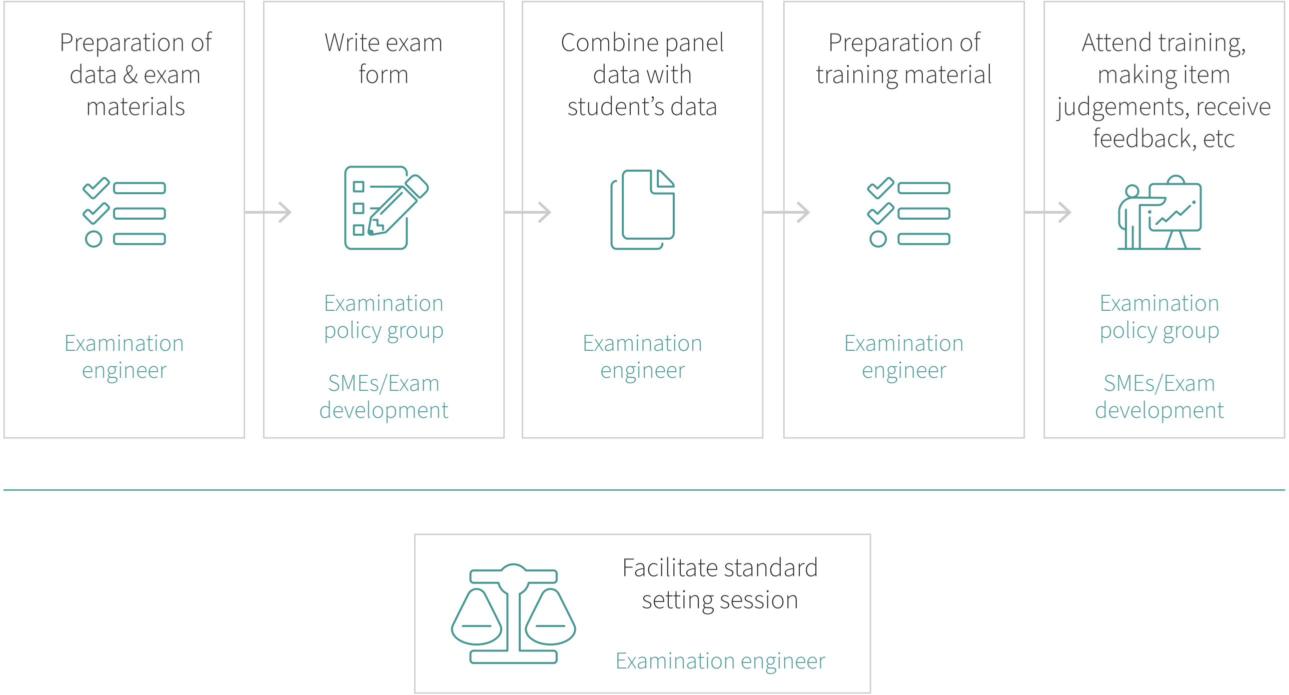
At this point, your new exam has been administered to candidates. Based on the pattern of candidate responses, each item has been reviewed to ensure that it meets psychometric standards. And from this analysis, you now have a well-vetted and validated set of candidate exam scores. What's left to do now?

Well, how many candidates actually passed your exam? And how did you make that decision? Although you have a list of candidate scores, you still have to decide on the score that candidates will need to achieve in order to demonstrate sufficient competence (and thus, pass your exam). Stage 7 of the Assessment Life Cycle is concerned with the process of establishing a well-informed pass mark (or cut score) for an exam. This cut score will be used to classify your candidate pool into two distinct categories: "competent" candidates (i.e., those who score at or above the cut score, and are deemed to possess the minimum level of proficiency required for safe practice); and, “not-yet-competent” candidates (i.e., those who score below the cut score, and are deemed to not possess the minimum level of proficiency required for safe practice).

The process used to decide on this cut score is called “standard setting”, and typically involves a panel of subject matter experts. The experts on this panel estimate the difficulty of each question in the examination. These item-level judgments are then averaged across all scored items in the examination, in order to derive a cut score for the examination as a whole.

An overview of the general process behind standard setting is provided below:

STANDARD SETTING



Preparation of data & exam materials
Examination engineer

Write exam form
Examination policy group
SMEs/Exam development

Combine panel data with student’s data
Examination engineer

Preparation of training material
Examination engineer

Attend training, making item judgements, receive feedback, etc
Examination policy group
SMEs/Exam development

Facilitate standard setting session
Examination engineer

1 | Selecting Standard Setting Panelists

In order to derive a valid cut score, it is important to be careful in your selection of standard setting panelists. As a group, the panel members must represent the breadth of the profession; and should bring a wide range of perspectives that constitute expertise in the domain being assessed. Each panelist will often have a very specific background in the domain being assessed on the exam; which allows them to offer well-informed opinions on how candidates will do on items that fall within that area of expertise. In addition, different panelists may be recruited from different key geographic regions (e.g., every province for a Canada-wide assessment); as well as from various professional practice contexts (for example, a health care certification exam may include standard setters from hospitals, community-based practices, and private practices).

The key characteristics of the standard setting panel should be clearly documented to ensure that the domain is well-covered, and to promote the defensibility of the process. By ensuring that the standard setting panel covers the breadth of the profession and examination's content domain, you are well-positioned to ensure that the standard that candidates are ultimately held to is informed by the collective wisdom of the profession.

2 | Conducting Pre-Meeting Preparations

Before actually conducting the standard setting meeting, some pre-meeting preparations need to be made. These include preparatory work on the part of both the standard setting panelists, as well as the psychometrician (or other test development expert) who will be facilitating the session.

PREPARING AS A SUBJECT MATTER EXPERT

Before the standard setting meeting, all panelists should familiarize themselves with the competency profile that the examination is based on.

If possible, it is also a good idea to have the panelists take the examination as a candidate would. Putting themselves in the position of actual candidates can prove to be a valuable perspective-taking opportunity. The experience of actually writing the exam often affords panelists a more realistic sense of the difficulty of each question than they would glean from just reviewing the question content. It is easier to facilitate this process when the examination is administered using computer-based testing (as the panelists will not need to travel to a specific location or manage sensitive materials like they would if the exam were exclusively paper-based).

PREPARING AS A PSYCHOMETRICIAN (OR OTHER FACILITATOR)

There are two primary ways in which the facilitator needs to prepare. First, training materials need to be created. These materials are designed to give panelists a clear sense of the task; and ensure that they have a sufficient and comprehensive understanding of standard setting prior to engaging in it. The specific training materials that are used may differ somewhat (according the specific standard setting method being used). However, some commonly-used materials include:

- A presentation that introduces the panelists to the standard setting process – and covers the background of the competency profile and examination blueprint.
- A short quiz on standard setting (or other comprehension check) to give panelists after the training.
- A list of behaviors and knowledge that characterize “unacceptable”, “just acceptable”, and “above and beyond” performance for each competency area in the competency profile, if available (this is used to help panelists reach a shared understanding of what constitutes “not-yet-competent”, “minimally-competent”, and “advanced competence” for each area being assessed).

Second, the facilitator should also compile and review the results of the item analyses conducted in Stage 5 (as well as historic candidate performance data for more established exams). This is important because most standard setting methods require panelists to consider the performance of previous candidates when rating the difficulty of items.

3 | Facilitating the Standard Setting Meeting

There are actually several different approaches to standard setting; and in the coming sections, we'll discuss the unique aspects of each. That said, there are also a number of common themes that run through most approaches to standard setting (which will be discussed here).
The first of these common threads is the use of the competency profile. Any standard setting should begin by reviewing the competency profile that the exam is geared towards. In doing so, the standard setting panelists will be reminded to keep the operational definition of each competency in mind when they rate items designed to assess those competencies.

In addition, the panelists should also have a solid understanding of who they're targeting the cut score for (that is, who the target candidate is). The cut score for credentialing examinations is meant to distinguish between candidates who have just the minimum knowledge and skills required for safe practice (i.e., the "minimally-competent"), from those who do not (i.e., the "not-yet-competent"). Although it may seem like an abstract concept, it is crucial to have a good understanding of minimal competence during any standard setting. If the panelists do not have a clear understanding of the specific factors that characterize a minimally-competent candidate (rather than a not-yet-competent, average, or advanced candidate), they risk gearing their standard setting ratings towards the wrong group of candidates (and, consequently, risk setting an invalid cut score).

As you might imagine, minimal competence can be tricky to operationally define and develop a shared understanding of. So, what can the standard setting panelists do to help build this understanding? Well, there are two main sources of information that they can draw from. First, the panelists should reflect on their own lived experience as practitioners. While working in the field, the panelists have almost certainly had direct contact with other practitioners, whose knowledge, skills, and abilities likely reflect varying degrees of competence across the key areas in the competency profile (after all, everyone has stories about "that one co-worker…"). Indeed, the panelists may even reflect back on their own current or previous level of competence in particular facets of the profession.

The second source of information that panelists have available is the list of specific competencies that candidates will be assessed on. By reflecting on the specific tasks that are subsumed under each competency area, panelists are often able to define key behaviours that would constitute "unacceptable", "just acceptable", or "above and beyond" performance. For example, mistakes that would risk a patient's safety (in a health care setting) would likely constitute unacceptable performance. Conversely, demonstrating out-of-scope knowledge or the ability to handle tasks that are usually completed by a manager would likely constitute "above and beyond" performance. By listing some of the specific behaviours that distinguish "unacceptable", "just acceptable", and "above and beyond" performance on the part of job incumbents, the panelists can effectively create a 'road map' for understanding what could reasonably be expected of (respectively) a "not-yet-competent", "minimally-competent", and "advanced" candidate. This list of behaviours is often enshrined as a list of "performance-level descriptors"; and may be used as part of the training package that future standard setting panelists receive.

Beyond these resources, it is also important to ensure that all standard setting panelists know how to make well-informed ratings. Often, you'll find that panelists rely on flawed heuristic strategies when deciding on their ratings. For example, a panelist might assume that every difficult item will lead to random candidate guessing (if the exam uses a standard four-option multiple-choice format, this will result in an Angoff rating of 0.25 for every difficult question). Alternatively, a panelist might assume that every candidate will correctly answer all easy items (resulting in Angoff ratings of 1.00 for every easy question). By having the panelists rate a few example items before beginning the standard setting in earnest, the facilitator will be in a good position to identify and correct these idiosyncrasies before the group invests a lot of time in rating the entire exam.

Once these training activities are complete, the standard setting panelists may begin the process of rating each item in the examination, leading to the establishment of a well-informed, defensible cut score.

4 | Some Approaches to Standard Setting

THE MODIFIED ANGOFF METHOD

When using the (modified) Angoff method, each panelist reads through the examination, and makes individual ratings for each item. These ratings (called "Angoff ratings") represent the percentage of minimally-competent candidates who they feel would answer that specific item correctly. For example, if a panelist feels that 70 percent of minimally-competent candidates would answer a given item correctly, they would make an Angoff rating of 0.70 for that item.

When making these ratings, the panelists need to read each question carefully; and then make judgments about the difficulty of the question. These judgements should consider both the 'structure' of the question (e.g., phrasing, effectiveness of distractor options), as well as the difficulty of the competency that is assessed.

In most cases, these judgments will be made over multiple rounds (which allows the panelists to refine and calibrate their ratings). One common approach is to have each panelist make an initial independent rating of each question in the exam. Afterwards, the facilitator compiles the panelists' responses; and examines the distribution of ratings for each item. The facilitator will then 'flag' any items that had highly discrepant ratings (traditionally, this is any item where the lowest Angoff rating and the highest Angoff rating differed by 0.25 or more). The facilitator will then guide a discussion on that item; and solicit feedback from the panelists about why they made the ratings that they did.

At this point, if there is historic candidate performance data available for the item, the facilitator will share this with the panelists. This way, the panelists can see the number of previous candidates who actually did answer the item correctly. The purpose of this “reality data” – and the discussion more broadly – is to give each panelist an opportunity to critically reflect on their rating for the flagged item. Following this discussion, each panelist is given a chance to make a revised rating of the item that is informed by the discussion and performance data.

The facilitator will then take these ‘Round 2’ Angoff ratings and calculate the average Angoff rating across all panelists and across all items. This value will become the recommended cut score for the exam as a whole.

Up until this point, we’ve discussed the process of making Angoff ratings specifically for multiple-choice items (where candidates can either get a question fully correct or fully incorrect). However, the Angoff approach can also be applied to exams where participants receive partial marks for questions (for example, if candidates are given a constructed response question that is scored out of five points). The process of applying the Angoff method to these items – called the “extended Angoff method” – is similar to the modified Angoff method. However, the ratings that the panelists make represent something a little different. In these cases, each panelist will make a rating of the percentage of minimally-competent candidates who they expect will receive each possible score on the item. So, for a constructed response item that is marked out of five points, each panelist would estimate the percentage of minimally-competent candidates who would receive 0/5, 1/5, 2/5, 3/5, 4/5, and 5/5. As with the modified Angoff method, this is often done in two rounds; with a chance for substantive discussion and the incorporation of previous performance data between the ‘Round 1’ and ‘Round 2’ ratings.

THE EBEL METHOD

An alternative approach to standard setting is the Ebel method. This approach is often used in health care certification programs that use objective structured clinical exams (OSCEs), or other role-play/performance-based assessments. OSCE examinations generally include a series of stations that candidates must complete. Candidates receive scores for each station based on whether or not they performed a series of requisite behaviours correctly (for example, one point for remembering to position the patient correctly before administering an injection; one point for correct placement of the needle, etc.).

The Ebel standard setting method diverges slightly from the Angoff method, in that ratings are no longer made for individual questions. Instead, the panelists follow a two-step procedure. First, each panelist makes a rating for each station in the OSCE to categorize how critical it is for safe performance at an entry-

to-practice level. For example, each panelist may be asked to rate each station on: 1) how difficult the tasks covered in the station are ("easy", "medium", or "difficult"); and, 2) how important/relevant the tasks covered in the station are for entry-to-practice candidates ("important", or "essential").
After each panelist makes an independent rating for each station, the facilitator guides a group discussion on these ratings in order to reach a group consensus on each station's difficulty and importance/relevance. Once this is done, there should be a good distribution of stations that cover every possible combination of difficulty and importance/relevance (e.g., "easy and essential" stations; "medium and important" stations, etc.).

Once all stations have been categorized this way, the panelists will then make another rating for each type of station (not each individual station!). This second rating represents the average score that the panelist feels a minimally-competent candidate would achieve on that station type (for example, the average score that minimally-competent candidates would be expected to achieve of the "easy and essential" stations). As raw scores may vary for each individual station, these ratings are generally made as percentages – for example, minimally-competent candidates may be expected to score a 60% on the "medium and important" stations.

As with the Angoff method, these ratings are often made in two rounds, with a chance to discuss discrepant ratings and review previous candidate performance on the stations between the 'Round 1' and 'Round 2' ratings.

THE BOOKMARK METHOD

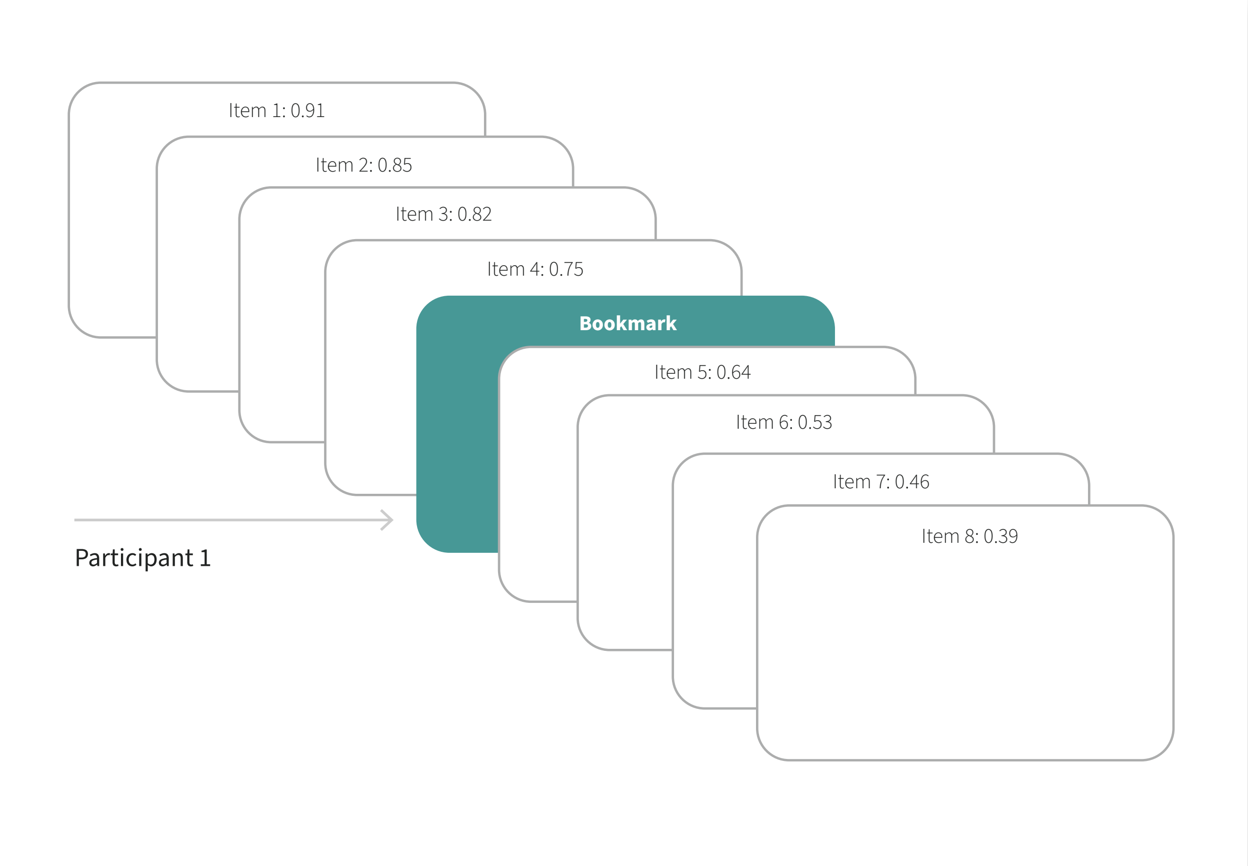
A third approach to standard setting is called the Bookmark method. This approach involves ordering all the questions on the exam form (or in entire the item bank) by their difficulty, from least difficult to most difficult. Question difficulty may be operationalized as their CTT p-value or IRT b-parameter (both covered in Stage 5 of the Assessment Life Cycle).

Once the items have been arranged by their difficulty, each panelist will then review each question. For every question, the panelist is asked to describe the knowledge that is required to answer the question correctly; and then reflect on what additional knowledge may be required to answer this question that makes it more difficult than the preceding question. After doing so, the panelist makes a judgment concerning whether or not they feel a minimally-competent candidate would answer the question correctly.

At some point, the panelist will reach an item that they feel minimally-competent candidates will not answer correctly. This item represents a "threshold" of sorts; one that separates minimally-competent from not-yet-competent candidates. The panelist places a "bookmark" between the item with the maximum difficulty that a minimally-competent candidate would answer correctly; and the item with the next highest difficulty (i.e., the first item that minimally-competent candidates would not answer correctly).

To illustrate this idea, please see below:

This process is initially completed independently by each panelist. Once each panelist has set their "bookmark", the facilitator then reviews the placement of everyone's "bookmarks", and flags discrepancies. These discrepancies are then calibrated between panelists through several rounds of discussion (which may also involve reviewing historic performance data). Through this process, the panel typically moves towards consensus on where the "bookmark" should be placed. This final "bookmark" is then used to inform the establishment of the examination's cut score.



Conclusion

In summary, the Assessment Life Cycle is a way of organizing the processes involved in creating valid assessments into a series of easy-to-understand, logical stages. The focus of this white paper was to detail the fundamental steps and key processes that are involved in the seventh of these stages (i.e., standard setting). Standard setting is the process of establishing a valid and defensible cut score for an examination using the input of subject matter experts (who represent the breadth and wisdom of the industry). There are several different approaches to standard setting, including the modified Angoff method, Ebel method, and Bookmark method.

Following these best practice steps – and the Assessment Life Cycle in general – will help ensure that your assessment program is valid and defensible; affording the greatest possible benefit to both your test-takers and your organization.

Let Meazure Learning help you apply the Assessment Life Cycle to your assessment program. Meazure Learning offers a full range of products and services that cover every step and process. Our clients agree: we know testing; and we will work hard to make sure that your testing program is the best that it can be.

To explore this opportunity – or for more information – please feel free to contact us at:
meazurelearning.com/services

List of Psychometric Services offered in Assessment Life Cycle Stage 7

At Meazure Learning, we provide a host of services to our clients that encompass each of the Assessment Life Cycle stages. Below is a list of psychometric services that Meazure Learning offers specifically for **Stage 7: Standard Setting:**

Service	Description
Standard setting facilitation	Standard setting is an important stage in the Assessment Life Cycle. The cut score you derive from the standard setting process serves to distinguish candidates who possess the minimum level of knowledge required for safe practice, from those who do not. Depending on the structure of your exam and the needs of your assessment program at large, there are several different approaches to standard setting (including the modified Angoff method, Ebel method, and Bookmark method). At Meazure Learning, we have a wealth of experience conducting standard setting sessions of various types and within several different organizational contexts. We'll work with our clients to decide on the best approach to standard setting; and facilitate the process of establishing a valid and defensible cut score.
Automated Angoff estimation services	With very large items banks, it can be financially-challenging to standard set every item. To help establish a valid cut score in these situations, Meazure Learning also offers automated Angoff estimation services. These use advanced computer algorithms to estimate Angoff values for items based on characteristics in the item metadata, as well as item-level statistics.