CU-101.9 - Introduction to Standard Setting

0:02
Welcome to Certiverse University and to this lesson, Intro to Standard Setting.

0:07
My name is Lance Blackstone, I'm one of the founders of Certiverse, and I'll be taking you through this lesson.

0:12
In this lesson, you'll be introduced to the concept of standard setting, that is, setting one or more levels of proficiency into which we will place test takers.

0:22
In the case of many exams, this is simply the cut score, or the score at which a test taker is considered to have passed the test.

0:31
We'll touch on the variety of methods available for standard setting, then go deeper into a common method known as modified Angoff.

0:39
Deciding the cut score or scores is a critically important step.

0:44
Setting it incorrectly has real significant consequences.

0:48
If you do all the other steps correctly, you can still end up with a poor exam.

0:52
If the standard isn't set well, people who should pass may end up failing or vice versa.

0:59
Even the perception that the standard has been set poorly can be damaging to your test and program.

1:05
Before we get started, quick disclaimer, this course is intended to be general and not Certiverse specific.

1:12
However, some industry concepts and terms are variable or can be ambiguous.

1:16
In those cases, we at Certiverse sometimes settled on a specific usage.

1:21
Thanks again for joining me.

1:23
Let's get started lesson.

1:25
We'll be covering the following Why do we even perform standard setting?

1:30
We'll talk about the goals and outcomes we expect.

1:34
Briefly, we'll cover several of the methods used to perform standard setting.

1:39
Then we'll dive deeper into a specific method that is broadly and commonly used in competency testing, the modified Angoff method.

1:47
Finally, there will be a bunch of relevant terminology you'll learn along the way.

1:52
Note that all terms are defined in a glossary you can find on the Certiverse Knowledge Base.

1:58
Standard setting is used to define levels of performance on an exam.

2:03
We define one or more distinct levels to indicate proficiency.

2:07
Keep in mind that the type of assessment will determine the levels needed, or that may even be possible.

2:13
For example, certification exams typically almost always simply care about competent or not competent.

2:21
This is often referred to as pass fail testing.

2:25
Another example might be a test for math skills targeted at grade 7 in US schools.

2:30
This test may need to stratify students into multiple levels like below basic, basic, proficient, and advanced.

2:39
Note the test itself must provide enough information to inform placing a test taker within a level.

2:46
For example, a math test like I just described would need to administer test questions across ability levels and enough of them to allow inferences about these levels.

2:58
On the other hand, a certification test will ideally cluster items at the specific single proficiency level we care about to strengthen our inferences at that level and only at that level.

3:12
This is why these types of tests should be careful about how they provide feedback beyond pass fail score.

3:21
For each proficiency level, we need to determine the quote cut score or the score that corresponds to that level of proficiency.

3:30
Typically, we use a candidate's raw score or points achieved for this purpose.

3:35
Referring back to our 7th grade math test as an example, we may determine that the level Basic equals 58 points achieved, Proficient equals 73, while Advanced equals 95.

3:48
That means that anything below 58 is below basic.

3:52
More relevant to most watching this video, for an exam used to determine minimal competencies such as a certification or licensure exam, we might say that passing equals 65 points.

4:04
This means that anything less than 65 points is failing.

4:08
Most importantly, the standard must distinguish those who have achieved the standard passed from those that have not failed.

4:19
These levels allow for interpretation of scores.

4:22
In the math example, we are placing students or learners on a continuum, often with the intent to provide diagnostic formative replacement feedback.

4:32
In the case of a competency exam, we are making the determination of competent or incompetent, which may be linked to certification or licensure requirements.

4:42
Of course, we need to be sure we are aligning the cut score to the intended purpose of the assessment.

4:49
We'll talk more about this in a bit.

4:53
There are a number of methods for performing standard setting, each with pros and cons.

4:58
More importantly, some methods are just better suited to specific types of testing than others, so it's important to use the method best suited to the specific exam.

5:08
The bookmark method is an empirically driven method that relies on item level statistics.

5:14
This means that the items are delivered to, ideally, a representative population of test takers.

5:21
Prior to standard setting, real administration data is used to calculate item difficulty using something called Item Response Theory or IRT.

5:31
Then, during standard setting, each SME places bookmarks in a real or virtual booklet where the items have been ordered by their difficulty.

5:42
The bookmark indicates the cut scores for each required level.

5:47
These bookmarks are aggregated across SM ES and presented back for discussion, which may result in updated ratings and ultimately A consensus or close enough is achieved in the PRO column.

6:00
The bookmark method utilizes empirical data on item difficulty.

6:04
However, this Pro is directly linked to a CON.

6:07
That is, this method requires prior test administration and a sufficient amount of data for accurate item calibration.

6:16
The contrasting groups method compares score distributions of groups known to differ in performance on an external criterion.

6:24
An external criterion is something outside the test that is ideally diagnostic of the groups taking the test.

6:31
For example, one might hypothesize that people with a valid driver's license should pass a written driving test, while those without should fail.

6:41
The external criterion

6:42
in this case is has or does not have a driver's license.

6:46
An advantage of this method is that it uses performance data to differentiate between groups.

6:52
A con is that the external criterion must actually be valid for grouping examinees.

6:58
Referring back to the previous example, it could turn out that holding a driver's license is not actually a valid criterion.

7:05
An inappropriate criterion criterion may not directly reflect the definition of competency being measured.

7:12
In many cases, there may just not be a valid external criterion, making this method not an option.

7:20
The borderline groups method analyzes scores of individuals identified as being on the threshold of passing or failing.

7:28
An advantage is that this method can incorporate nuanced and contextual information about test takers.

7:34
A disadvantage is that identifying truly borderline individuals can be very challenging, and not having enough of them could result in validity issues.

7:45
Finally, in the Angoff method, experts review each item on an exam form and estimate the probability of a minimally qualified candidate answering each item correctly.

7:56
This data is analyzed and where the differences exist, discussion occurs.

8:00
In some cases, empirical data is used as a reality check to better align SME ratings to actual test taker data.

8:07
Ultimately, final ratings are average to calculate the cut score for each level.

8:13
An advantage of this method is that it can be used before test administration.

8:19
Unless, of course, you're using administration data as part of the process, and it's relatively straightforward to understand and implement in the con column.

8:29
This method relies heavily on expert judgement, which can be subjective.

8:34
There's the potential for overconfidence bias.

8:37
To overcome this, understanding each level of proficiency is critical.

8:41
In the case of competency testing, we lean heavily on the definition of the minimally qualified candidate for this understanding.

8:48
In the modified Angoff method, we may also introduce empirical data in the form of test taker performance to further ground our decision making around the standard being set.

9:00
Whatever the method used, the purpose of any standard setting is to use an accepted, transparent, replicable methodology to make a subjective judgment that is as objective as possible.

9:15
Let's drill down on one of the most common methods used for competency testing, the modified Angoff method.

9:21
Here are the major phases of this process.

9:25
First, we need the definition of each level, typically simply the minimally qualified candidate.

9:32
If we've performed a JTA, we should already have this.

9:35
We also need a group of SM ES that will be engaged in this effort.

9:39
We call this the Standard Setting Panel or just the panel.

9:43
We then select a version of the exam or a form to use.

9:48
This form contains all of the items that will be rated.

9:52
In fact, the Standard Setting form does not need to be a real form at all.

9:57
That is, it may only exist for the purpose of standard setting.

10:00
Further, it may contain all of the items to be used across all the forms.

10:06
This can be a real advantage in terms of computing cut scores for each form simply based on the items that end up on that form individually.

10:15
SMEs will rate and sometimes rewrite each item on the standard setting form and estimate the probability of a minimally qualified candidate answering each item correctly.

10:27
In some cases, empirical data is used as a reality check to better align SME ratings to actual test taker data.

10:35
The standard setting admin or facilitator will take the SM ES ratings and analyze them.

10:41
This includes aggregating ratings as well as identifying notable discrepancies across raters and items.

10:48
A series of meetings are needed, initially to align on goals and the process, and then later, once item evaluations are available, to discuss the ratings and address discrepancies.

11:00
After all of the above, the cut score or cut scores where multiple levels are being established, is set.

11:08
Often the panel makes a recommendation to the certifying body that in turn sets the final cut score or standard.

11:15
Finally, documentation is created or finalized which will support the validity and legal defensibility of your exam.

11:25
We need to clearly define the expectations for each level that we plan to designate.

11:30
Minimally,

11:31
our exam will have two levels, but it might have more.

11:34
We need to clearly define what competency at each level looks like.

11:39
When we are talking about exams leading to certification or licensure, again, we are generally making a binary decision.

11:46
Competent or not competent, pass or fail.

11:49

We refer to this definition as the minimally qualified candidate, and test takers either meet that standard or they do not.

11:58
As a reminder, the MQC is not a real person.

12:01
It is a representation of the theoretical person who's just qualified, barely qualified to do the job, not overqualified, not highly qualified, minimally qualified to enter the job or role.

12:16
This concept underpins all the follow-on decisions to be made during standard setting.

12:20
Your SM ES must understand and agree on who the minimally qualified candidate is and apply this understanding to setting the passing standard.

12:29
The panel is the team of individuals who will be tasked with evaluating items in order to set the standard.

12:38
This group may range in size from 6 to 20 individual subject matter experts.

12:43
They may be the same, overlapping, or a completely different group from those SM ES involved in previous steps such as the JTA or blueprinting.

12:53
The team should be as representative as possible.

12:56
Different exams may have differing requirements when it comes to finding a representative panel, but standard demographics are often relevant, such as physical location of practice, gender and age.

13:08
Other considerations may be specific to your exam.

13:11
For example, an exam about databases may need to include database developers, application developers and database administrators.

13:19
An exam about commercial food handling may need to include chefs, line cooks, management and front of house staff.

13:27

Once assembled, this group needs to be trained on the processes and the expectations for their contribution.

13:34
The panel needs to clearly understand the definition of the MQC and agree with it.

13:39
This is very important.

13:43
They will need to understand the follow-on processes as well how they are expected to review items in subsequent meetings to focus and sometimes resolve discrepancies.

13:55
In order to do standard setting, first we need a form of the exam to work with.

14:01
Again, a form is simply a specific version of the exam.

14:06
Often in high stakes testing, multiple forms exist to increase the integrity and security of the exam.

14:11
For example, multiple forms allow test takers to retake the exam while ensuring they don't benefit from previous attempts as they will see a mostly or completely different set of items.

14:21
Multiple forms also provide some level of protection for malicious actors attempting to memorize or record items to share with others.

14:30
So the standard setting form is simply the form chosen to use for standard setting.

14:36
If multiple forms already exist, then the choice of form to use for this purpose can be arbitrary.

14:42
However, as mentioned before, another option is to create a special standard setting form that includes all of the existing valid items.

14:51
Performing standard setting across all items allows for computing a cut score for any forms made from those items.

15:01
Once impaneled, your SM ES will be asked to rate each item on the standard setting form.

15:07
To enable this, each panel member is provided the items either physically or electronically.

15:13
This work is done by each of your SM ES alone.

15:16
We want their initial read of each item without any discussion.

15:22
SM ES are asked to provide for every item on the standard setting form their estimate of the proportion or percentage of minimally qualified candidates that they think will get the item fully correct.

15:36
This means that the higher the rating, the easier the item.

15:40
An item rated at 50% means that the SME thinks that only half the minimally qualified candidates will answer that item correctly.

15:48
This is a pretty hard item.

15:50
In comparison, an item rated at 95% is considered very easy.

16:00
Once all SMEs have rated all items, the facilitator reviews and analyzes the data.

16:06
Using the SME ratings data, the facilitator will calculate average ratings.

16:11
The facilitator looks at the ratings across items to see where raters converge and more importantly, where they diverge.

16:18
The divergent ratings are the ones that will need to be addressed in subsequent meetings.

16:23
In preparation for that meeting, the facilitator will need to pull together their analysis focused on areas of disagreement.

16:29
They may also pull an additional data points, such as actual response data if possible.

16:34
In this case, the relevant data is the calculated difficulty of the item from real test takers.

16:40
They may also provide calculated stats like inter rater reliability.

16:45
A statistical evaluation of the degree of agreement among panel members.

16:54
The panel discussion step is the first time your SM ES will be able to discuss the ratings with other SM ES.

17:01
The panel is provided with the outcomes of the previous analysis step.

17:05
This will include some or all of the following aggregated statistical values for item ratings, individual SME item ratings, additional data such as actual item performance stats if available, inter rater reliability statistics, and of course highlighting of major discrepancies.

17:27
The facilitator then guides the conversation using major discrepancies to unpack how SM ES came to their ratings.

17:34
Discrepancies often arise out of differences in understanding of the minimally qualified candidate definition, differences in how other evaluation criteria have been applied, and differences in how individual SM ES might read and understand the item itself.

17:55
Consensus is not required, but understanding discrepancies can highlight issues with items and help resolve them.

18:02
The discussion can unpack differing viewpoints, so here's an example.

18:08
Let's say one SME rates an item as very hard, while another rates it as very easy.

18:15
Through discussion, we may uncover that the first SME sees the item as an esoteric piece of information not relevant to a minimally qualified candidate.

18:25
A second SME might have rated the item very easy though.

18:29
During discussion you find out that the second rater actually agrees that the item is esoteric.

18:34
However, the rater goes on to point out that the distracters are not good and that the item, while esoteric, is still simple or easy to answer using process of elimination.

18:48
This discussion of the rating discrepancy gives us insight into how the two SM ES perceived the same item as both hard and easy.

18:56
This allows for subsequent decisions about that item.

19:01
The process of individual item rating and meeting may be repeated multiple times.

19:06
Some processes attempt to converge on a consensus.

19:09
For example, the process may require that all raters be within 20% of each other on each item.

19:15
That is, I can rate an item at 40% while you rate it at 60%.

19:21
On the other hand, some processes don't worry so much about consensus and simply rely on statistical averaging of SME ratings.

19:28
These tend to focus more on the benefits achieved by identifying problematic items, the ones with large discrepancies.

19:35
There can certainly be too much emphasis placed on consensus, and there can be dramatically diminishing returns from too much discussion.

19:45
So we're almost done, but not quite.

19:48
Once we have the outputs from the previous steps, we can work towards a final cut score.

19:54
1st, we average panel ratings to set the expected percentage correct for each item.

20:01
If multiple rounds of ratings happened, we use the latest ratings.

20:06
We can now easily calculate the expected cut score by taking the expected percentage correct for each item, adding them all together, then dividing this number by the number of panel members.

20:20
We might be done at this point, however, some programs will go further if there is time, budget, and data.

20:27
Here are some examples.

20:29
What are the calculated min and Max cut scores?

20:33
That is, if we take each individual SM ES ratings, where would they set the cut score based on their own individual rating?

20:43
Are the lowest and highest values here reasonably consistent with the average cut score?

20:50
The next technique, called the Beuk Compromise, requires that we have real test taker data.

20:56
Assuming we do, we can ask each SME to estimate the pass rate at the calculated expected cut score.

21:04
We then compare these estimates to the actual pass rate at that score.

21:10
I'm going to avoid going into too much detail here, but the idea is to provide another reality check.

21:16
If your SM ES think that 70% of test takers should pass at a given cut score, when in really reality only 25% pass at that score, you might have a problem.

21:29
A final example is the use of standard error of judgment as an element in adjusting the cut score.

21:35
Standard error of judgment is the average amount of error or variability in judgments made by a panel of judges or raters.

21:43
That's a bit of a mouthful.

21:45
It essentially quantifies how much the individual judgments tend to differ from the overall average judgment.

21:53
This can be used to adjust the cut scores down in the candidates' favor.

21:58
To account for this error, the process of standard setting decisions and outcomes need to be documented.

22:06
Doing so aligns to best practices, provides validity support for the exam, and can be used to legally defend your program if needed.

22:16
It's really best to continuously document the process as if assembling the final documentation and report will be easier and more accurate.

22:25
Software automating this process is highly beneficial.

22:30
Standard setting reports are often shared publicly.

22:33
When we certify or license people, it is very reasonable for the public to understand what is involved.

22:39
Transparency is important.

22:43
The report itself will generally contain the following, a description of purpose that is consumable by all stakeholders, including the general public.

22:54
The methodology needs to be discussed, including your specific implementation.

23:00
The individuals contributing to the standard setting should be listed, including their credentials.

23:07
The rounds of review, discussion notes, data generated, and any outputs should also be captured.

23:16
The final cut score recommendation should be reported along with any and all additional considerations that resulted in modifications to that cut score.

23:28
There's a whole lot more we could talk about here, but this is hopefully a good conceptual introduction to both standard setting in general and the modified Angoff approach specifically.

23:40
I hope you enjoyed this lesson and I hope to see you in the next.

23:44
Till then.