Intro to Item Analysis with Alan Mead

0:02
Welcome to Certiverse University, and welcome to this video Introduction to Item Analysis.

0:07
My name is Alan Mead and I'm the Chief psychometrician for Certiverse.

0:10
In this video, we'll cover how to analyze the data that you just collected in the beta test and how to interpret those results to improve your exam.

0:17
Let's get started.

0:22
Once you have your item level data from your beta exam, you perform item analysis, which determines whether each item is doing its job well or poorly.

0:30
I like to use a stoplight model.

0:32
Then the goal of the analysis is to classify items into red, yellow, or green categories so that you can eliminate the red items which are thwarting the purpose of your exam, actually harming it, and limit the number of weak yellow items.

0:46
So item analysis is just a matter of classifying items into the stoplight model.

0:50
And by the way, I'm going to talk about item analysis using classical test theory, also called CTT, which is a simpler approach.

0:58
But if you understand that type of item analysis, you can do exactly the same thing with more advanced analysis like item response theory or Rasch.

1:11
The two key metrics are item difficulty and item discrimination.

1:17
Item difficulty is just the proportion of candidates who get an item correct.

1:20

This can be called AP value, although that's ambiguous and I would discourage it because there's another statistic called P value that means something completely different.

1:28
It can also be called the proportion correct and abbreviated PC, and it's also the average of candidate scores on a scored item.

1:36
So it's sometimes labeled as the mean because mean means average, and you may even hear it referred to as facility, which means easiness as in facile because the scale is backwards.

1:48
So difficulties below .5 mean that less than 50% of your sample is getting the item correct.

1:54
Those are hard items with difficulties close to chance of responding being extremely difficult.

2:04
Difficulties above .85 or so, meaning that 85% or more of your sample are getting the item right, are easy, with difficulties close to one point O being extremely easy.

2:14
Because everybody or almost everybody's getting the item right.

2:17
It's hard to say what an improper difficulty is for an individual item.

2:21
Items should vary.

2:23
There should be some easy and some hard and a lot of medium difficulty items.

2:27
But being extremely difficult may be an indication of a problem, particularly when the difficulty is below the chance level.

2:32
And for a four option multiple choice item .25 is chance level responding.

2:37
So you might ask, why calculate item difficulty?

2:40
Well, it's useful when you look at forms.

2:41
For example, let's say Form A is too difficult.

2:44
You could sort the items by difficulty and swap some of the most difficult Form A items with some of the easier Form B items to fix this issue.

2:56
Discrimination is how well individual items distinguish between candidates who should not pass and those who should pass.

3:04
In most cases, discrimination is a bad thing, but here in this context, it's a good thing.

3:08
And discrimination is operationalized as a correlation between the scored item and the total exam score.

3:15
So it's called the item Total correlation or the ITC.

3:18
And a correlation is a decimal number, So a strong positive ITC is good, a strong negative ITC is bad.

3:25
And Itcs near zero are not great, but they aren't actually harming your exam score reliability.

3:31
They're taking up space without contributing much.

3:34
The corrected item total correlation, or the CITC, is the correlation of the item with the rest of the items on the exam.

3:42
And I think the CITC is a better indication of item quality.

3:47
So I'm going to recommend that we use that.

3:52
Here's a chart showing one way to classify items into the stoplight model.

3:57
An easy item is always yellow.

3:58
A strong negative citc usually makes an item red.

4:02
But if almost everyone gets the item correct, the citc is influenced by the small number of people who got the item wrong.

4:09
So in that case it's yellow.

4:10
Items that are very difficult are yellow, or if they're close to being at below chance level, red.

4:15
Here we're assuming four option multiple choice items to calculate chance level responding, which is .25.

4:21
Other items are green, including those that are at least adequate.

4:24
Maybe the details of this are fuzzy, but two key points are that there are a lot of ways that an item can be imperfect and be red or yellow, but most of the time you'll have a small number of red items and mostly yellow and green items, hopefully the majority of them green.

4:41
This, by the way, is the reason that you need 50 to 100 items on an exam to get reliable results.

4:47
Most exam items are at least a little bit flawed, but when you put enough of them together, you get a reliable score.

4:58
Now, as I said, we've been discussing item analysis using classical test theory, or CTT.

5:03
Modern psychometric theories like IRT and Rasch are more complex, but they have statistics that can be used equivalently.

5:10
Here are some of the differences.

5:12
IRT models require non negative discriminations, so some red items we dropped during the analysis,

5:19
you won't even see the item in the analysis because it has to be dropped.

5:22
Also, IRT models can model degree of guessing and too much guessing is bad.

5:28
That means that the probability that a person can get the item right by chance alone is unacceptably high, so they have that additional parameter.

5:40
Rasch models assume that the discrimination for all items is the same, so any item that is better than the average or worse than the average gets dropped from the analysis.

5:51
This means that some red items are dropped, but also some green items.

5:55
Also, Rasch assumes no guessing, so guessing could damage the neat mathematical properties of the Rasch model, and this may cause problems with the analysis of multiple choice items.

6:05
Both of these points might make it seem like Rasch is a bad choice, but I think that misses a point.

6:11
Any analysis is a lot better than no analysis, and Rasch is widely used and successfully used.

6:18
All that said, item classifications using IRT models generally agree very closely with those using classical test theory.

6:29
So in the next slides I'll be showing you how to interpret an item analysis.

6:33
But that begs the question, how do you get the item analysis?

6:36
It is beyond the scope of what we can talk about to discuss the actual analytic steps.

6:41
But here's some ways that a client might obtain the statistics from their beta exam data.

6:46
1 is that platforms like Certiverse, and not just Certiverse but many platforms, will generate these reports automatically.

6:52
Or you might need to engage a psychometrician, give them your data and have them use statistical software to generate these reports.

7:00
Or it's possible that somebody in your IT department, a developer, may be able to produce these reports for you.

7:05
They can actually be done, or largely done using Excel, so they're not super complicated from a analytic perspective.

7:16
So here's a typical item analysis tabulation.

7:19
I know this is a lot of information, but I think you'll find it easier to see after I explain it.

7:24
First off, we only need to look at the left most columns.

7:27
The columns to the right show an option analysis.

7:30
We'll go over option analysis in the next video.

7:32
You use option analysis for a different purpose.

7:35
It's to guide the SMEs when reviewing and fixing flawed items.

7:39
Our purpose today is to identify those flawed items so we can ignore the option analysis and focus on the left hand side.

7:44
So on the left hand side we see columns for the item name, the item scoring status, the sample size labeled N, the corrected item total correlation labeled CITC, and the item mean, which is the difficulty.

7:59
That's the proportion of candidates who answered the item correctly.

8:02
Could also be labeled PC for proportion correct or P value or difficulty, or maybe even facility.

8:09
For this example, I focused on the essential columns.

8:13
In a real item analysis, you might have more metadata about the item.

8:16
For example, the Certiverse item analysis includes the item type, the content area, and the key or keys.

8:27
Before we start talking about interpreting these statistics, let's talk about sampling error.

8:33
And this is a deeply technical issue.

8:34
So the two key qualitative issues are, first off, that there is a degree of error in our statistics, and 2nd, that the error is smaller but not trivial for big samples, and it can be quite large for small samples.

8:48
So here's the technical bit.

8:49
When you see a sample statistic like a CIT C of .4, that's not the true population value because of sampling error.

8:57
For example, if we gathered a brand new sample of 275 individuals, we would be 95% confident that the CIT C would be as low as .28 and as high as .52.

9:08
That's a pretty big range for a correlation like the CITC.

9:13
The margin of error is complicated, but it's roughly twice the reciprocal of the square root of the sample size.

9:23
So if we take sqrt 275, that's about 16.6.

9:28
Pressing the reciprocal button gives us .06, and doubling that gives us .12.

9:34
For a sample size of 53, the margin is about .27.

9:39
Applying this to item 1, the CITC of .4 is ±.12, which makes us 95% confident that the true value is between .28 and .52.

9:51
And for item 13, the CITC of -.24 is ±.27, which makes us 95% confident that the true value is between -.5 and 0.

10:03
These are reasonably large ranges, and the range for a sample size of 275 is already bigger than we'd like, and the range for a sample size of 53 is quite large.

10:21
The margin of error for a mean is calculated differently.

10:24
I'll spare you some of the details, but the error is smaller compared to that of a correlation.

10:28
Typically, about half these items all have margin error around .05, including item 13.

10:34
Applying this to item one, the mean of .42 is plus or minus .06, which makes us 95% confident that the true value of between .36 and .48.

10:45
As for item 13, the mean of .96 is plus or minus point O 5, which makes us 95% confident that the true value is between .91 and one.

10:56
These are still uncertain values, but we have a greater degree of certainty about the outcome of these statistics in the true population.

11:10
So again, if all this reciprocal of the whatsis is confusing, here are the key takeaways.

11:15
Statistics have a margin of error.

11:17
We would expect to see at least slightly different values if we took a new sample.

11:22
This is why psychometricians keep asking you for what might seem, from a practical and logistics perspective, to be unreasonably large samples.

11:30
When you analyze small samples less than 100, the error might be substantial.

11:35
The error is larger for a correlation like CITC than it is for a mean.

11:41
So in tiny samples, I might not trust the CITC values at all and only focus on the mean difficulties, which are less informative at the item level about how good an item is doing.

11:57
So let's go through these items.

11:59
Item 1 is a scored item with a modest sample size of 275.

12:03
It is very strongly discriminating with the CITC of .4.

12:07
That's very good, but it is quite difficult with only about 42% of candidates answering correctly.

12:12
We would classify this as a yellow item, but only because of the difficulty and I would want to include this item on the exam unless there was a better alternative or the whole exam was too difficult.

12:22
Item 5 is a pretest item.

12:24
It has an adequate discrimination with the CITC of .22 and it's easy.

12:28
82% of the candidates got the item correct.

12:31
This item would be classified as green and it would be a fine item to include on an exam form.

12:40
Item 13 is a pretest item with a tiny sample size of 53.

12:44
Maybe we should pretest this item some more before I interpret this analysis.

12:48
But look at that terrible CITC of -.24.

12:51
That would be really bad.

12:52
But in the next column we see that the item is extremely easy.

12:56
96% of candidates answer this item correctly.

12:58
If you do the math, that means two of the 53 candidates answered incorrectly.

13:03
Those two candidates are driving the negative CITC.

13:06
They probably did well in the exam.

13:08
I would discount that CITC in this situation where the item is very easy.

13:13
A bigger issue is that almost everyone is getting this item correct.

13:16
It's taking up space in the exam without contributing much.

13:18
That's the bottom line for this item.

13:20
It's too easy.

13:21
We would classify this item as yellow.

13:23
I would include it if needed, but hoping for a better alternative to this item.

13:31
Finally, item 15 is a scored item and it looks bad.

13:35
The sample size is not tiny, the CITC is strongly negative, and the item's very difficult.

13:41
This is bad.

13:42
I would strongly suspect that this item is either too tricky or mis keyed.

13:46
We'll return to this item in the next video.

13:49
When we look at option analysis, we would classify this item as red and including it on a form would slightly damage the reliability of that form, so I would avoid it unless there were no other alternative.

14:02
I hope you've enjoyed this video and now have better insight into interpreting item analysis, which are a critical step in developing fair and reliable exams.

14:10
In the next video is an optional step.

14:13
If you want to try and salvage any red items, an option analysis will help your smeeze.

14:18
We'll cover option analysis in the next video.

14:21
After that, we'll cover additional steps in finishing the exam.

14:26
That's it for this video.

14:27
Thanks for watching.

14:28
We'll see you in the next.