STAGE 5

Item and Test Analysis

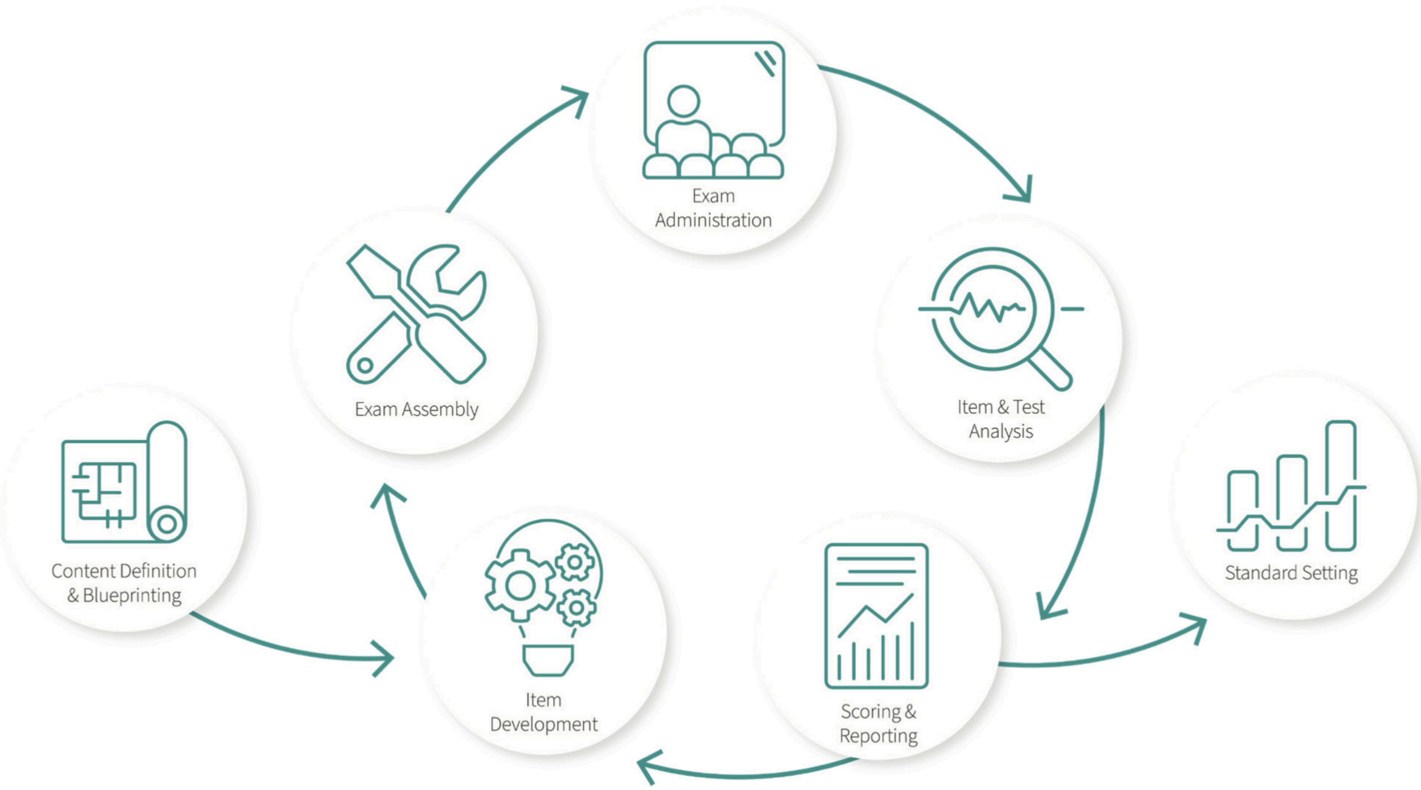
MEAZURE LEARNING

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 stage of the cycle, you have finished designing and validating your exam; and have presented the approved exam form (or forms) to your candidates. The exam has been written; and you are sitting on a pile of candidate response data. But your work is not over quite yet. Stage Five of the Assessment Life Cycle involves analyzing candidates' responses; and looking for patterns in their responses to evaluate whether the items in your exam have indeed created a psychometrically-valid assessment of candidates' ability.

To make this judgement, each item must be evaluated to ensure that it meets minimally-acceptable psychometric standards. Items that do not meet these standards are flagged for review; and may be revised before being included in future exams (or retired from the item bank altogether).

The purpose of this white paper is to explain how one conducts the analyses used to make these judgements – and the implications they have for future item development. As with other stages, there are actually several distinct tasks that are subsumed under the umbrella of item and test analysis. These include:

- Establishing standards for item quality and item reliability
- Data export and cleaning
- Code preparation
- Item and test analysis reporting
- Review and interpretation of analyses

1 | Refining Item Development Targets: Establishing Standards for Item Quality and Item Reliability

Before any actual analyses take place, the examination policy group must decide on the standards that they'll hold the items (and the exam) to. If an item meets these standards, then candidates' scores on that item can be assumed to genuinely represent the candidate's ability in the subject area being assessed. This decision must be made carefully; and should consider best psychometric practices, as well as any idiosyncrasies in the exam that may need to be taken into account (e.g., having a small sample of candidates, having an unusual candidate cohort). In order to help facilitate this process, a psychometrician is normally involved.

STATISTICAL STANDARDS FOR ITEMS (DIFFICULTY AND DISCRIMINATION):

When deciding on these standards, the examination policy group will normally consider two key factors: 1) psychometric/industry best practices; and, 2) the context of the testing program.

Perhaps the most common framework for understanding psychometric best practices is drawn from Classical Test Theory (CTT). Under CTT, items can be assumed to differ along two main continuums: difficulty and discrimination. Consequently, there are two (statistical) standards that are normally considered when conducting a CTT-based item analysis:

- First, items shouldn't be too hard or too easy. Ideally, all items should fall within a specific difficulty range. This is generally assessed using a p-value (i.e., the proportion of candidates who answered a given item correctly). For example, any item that more than 85 percent of candidates answered correctly (i.e., $p > .85$) might be flagged as being overly easy; whereas any item that less than 30 percent of candidates answered correctly (i.e., $p < .30$) might be flagged as being overly difficult.
- Second, candidates with higher ability should be more likely to answer any given item correctly. That is, items should be able to discriminate between high-ability candidates and low-ability candidates. This is usually assessed by examining the correlation between candidates' scores on any given item; and their overall exam scores (e.g., a corrected point biserial correlation, or CRPB). For example, the exam policy group may decide that the correlation between candidate scores on any given item should have at least a correlation of .25 with candidates' overall exam score (i.e., CRPB >= .25).
- When conducting an item analysis, these CTT indices are sometime supplemented with conceptually-similar indices drawn from Item Response Theory (IRT).
  When approaching the same two questions from an IRT-based perspective, a psychometrician may examine each item's a-parameter, to ensure that it meets or exceeds a base threshold (evidencing a minimal ability to discriminate between high-ability and low-ability candidates); as well as each item's b-parameter, to ensure that it that falls within a desired range (evidencing that the item is not too easy or too difficult). These indices will be described in more detail later in this paper.

STATISTICAL STANDARDS FOR EXAMS (RELIABILITY):

In addition to examining the performance of specific items, the exam policy group may also consider the overall performance of the items. This is usually done by examining the reliability of candidates' scores on the exam. Reliability is a measure of consistency in responses; and in a testing context, measures the extent to which scores on an examination are consistent with one another (that is, are candidate responding to each item on the exam in a similar way?).

Test reliability statistics are normally calculated using an index of the examination scores' internal consistency. Internal consistency provides an index of how closely candidate scores on each item correlate with candidate scores on all other items in the exam. Two commonly-used indices of internal consistency are the KR-20 (Kuder & Richardson, formula #20) and Cronbach's Alpha.

Of these, the KR-20 is used exclusively when examination scores are dichotomous - that is, the candidate can either answer an item correctly or incorrectly (as is the case for most multiple-choice exams). Conversely, Cronbach's Alpha can also be applied to situations when a candidate can receive partial credit for a question (for example, a constructed response item where the answer is scored out of five points). Although useful for partial credit questions, Cronbach's Alpha can also be used when dealing with dichotomously-scored items; and when candidate answers are dichotomous (that is, all items are scored as either right or wrong), the KR-20 and Cronbach's Alpha will yield identical results. In these cases, a commonly-used used standard for certification/licensure exams is to ensure that the KR-20 or Cronbach's Alpha exceeds .90 (indicating a high level of internal consistency in candidate responses).

So what factors improve (or diminish) reliability? Well, there are actually several factors that can affect the reliability of a set of exam scores, including:

- **Item difficulty:** Having questions that are either very difficult or very easy will limit the ability of these items to effectively discriminate between high- and low-ability candidates. Consequently, the reliability of these candidates' test scores will be negatively impacted.
- **Item discrimination:** Similarly, questions with high discrimination indices (e.g., CRPBs) will be more effective at differentiating candidates with high- versus low-ability. This will often improve the reliability of candidates' test scores.
- **The construct (or ability) being measured:** If all of the questions on the test are meant to assess the same underlying construct (for example, an exam that exclusively tests candidates' clinical nursing skill), then candidates scores across all items should

theoretically be more similar than if they were measuring different constructs - thus increasing the overall reliability of test scores. Conversely, an exam that assesses multiple theoretically-unrelated constructs (for example, an exam that tests candidates' clinical nursing skill and their knowledge of jurisprudence), will likely have more heterogenous test scores across candidates; leading to lower score reliability.

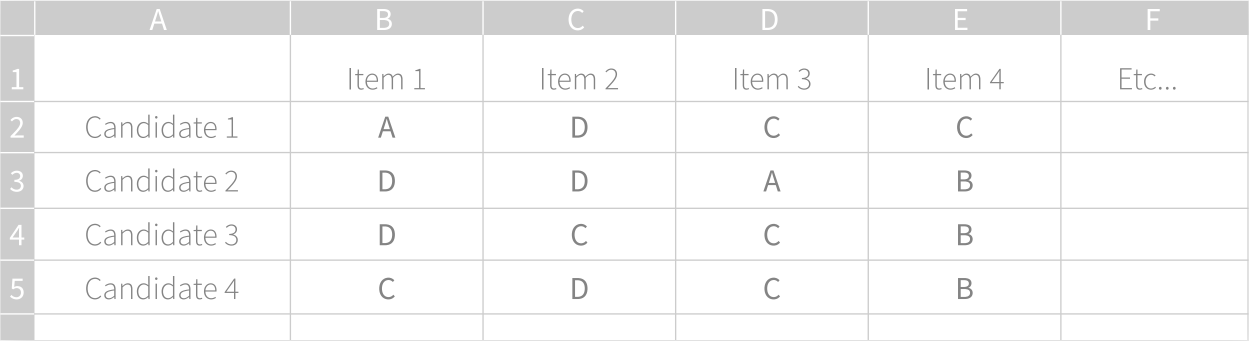
- **Sample size:** Candidate scores are more robust when you have a lot of them. When you only have a few candidates in your sample, the mean of their scores will tend to bounce around (as a single outlier will have a greater impact on the overall mean). Consequently, smaller samples of candidates often have lower reliability indices.
- **Number of questions:** The more items you have on your exam, the higher the reliability of scores will generally be. However, there are two caveats to keep in mind. First, having more items only improves reliability if the items you're adding are robust and valid (and thus, accurately measure the subject matter they're supposed to). Second, adding more items only tends to improve score reliability to a point; and you will often see diminishing returns as the exam becomes larger. This may be for the best – we wouldn't want to have an exam form with 5,000 questions after all!
- **Extraneous variables:** Many factors can interfere with a candidate's ability to answer a test question correctly; and these random "nuisance" factors can often hurt the overall reliability of test scores. For example, imagine that halfway through an exam, loud noises outside the test centre begin to distract the candidates (causing them to perform more poorly on the second half of the test). Any resulting item analysis will suggest that scores on the first and second halves of the exam are inconsistent (even if this is due to an extraneous factor rather than the items), decreasing the reliability of the scores.

2 | Data Export and Cleaning

Once the examination policy committee has made well-informed decisions concerning the statistical standards that items and exam forms will be held to, it's time to export and clean the data. This is a necessary step that will help set the stage for the actual item analysis.

To export data, one normally downloads candidates' raw scores from the examination platform; and then stores them in a data file (for example, a series of comma-separated values, or .CSV file). This data file can then be accessed and analyzed using a data analysis program such as Excel, SPSS, or R.

Usually, each row in the data file represents a single candidate; whereas each column represents the response to a given item (see below for an example of this set-up).



	A	B	C	D	E	F
1		Item 1	Item 2	Item 3	Item 4	Etc...
2	Candidate 1	A	D	C	C	
3	Candidate 2	D	D	A	B	
4	Candidate 3	D	C	C	B	
5	Candidate 4	C	D	C	B	

Once the data has been successfully exported and loaded into the data analysis program of your choice, you can begin the process of data cleaning. Data cleaning involves reviewing the data for any inconsistencies and outliers (i.e., extreme values). This data cleaning actually includes many different types of integrity checks. These might include:

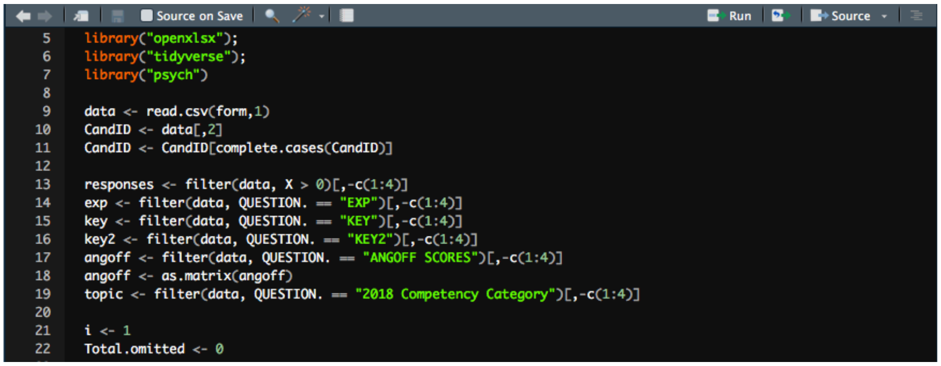
- Ensuring the data range for responses is correct
- Ensuring that the data file contains the correct number of questions
- Ensuring that the data file contains the correct number of candidates
- Ensuring that any non-scored (i.e., experimental) items are coded as such, so that they won't be counted towards candidates' scores
- Ensuring the questions are coded to the correct examination blueprint category
- Checking to make sure there isn't an excessive number of non-response cases

Once the integrity of the data file has been confirmed through this cleaning process, it is ready to be analyzed!

3 | Code Preparation

When conducting an item analysis, it is critical to be transparent about what it is you're doing. That is, another psychometrician (or other testing expert) should be able to understand and replicate your analyses.

Thankfully, many data analysis platforms make it easy to ensure this transparency by allowing you to save your analysis "code". For example, platforms like SPSS or R allow you to save a syntax file (i.e., a set of commands written in a kind of programming language). Consider the example below:



```
library("openxlsx");
library("tidyverse");
library("psych")

data <- read.csv(form,1)
CandID <- data[,2]
CandID <- CandID[complete.cases(CandID)]

responses <- filter(data, X > 0)[,-c(1:4)]
exp <- filter(data, QUESTION. == "EXP")[,-c(1:4)]
key <- filter(data, QUESTION. == "KEY")[,-c(1:4)]
key2 <- filter(data, QUESTION. == "KEY2")[,-c(1:4)]
angoff <- filter(data, QUESTION. == "ANGOFF SCORES")[,-c(1:4)]
angoff <- as.matrix(angoff)
topic <- filter(data, QUESTION. == "2018 Competency Category")[,-c(1:4)]

i <- 1
Total.omitted <- 0
```

Writing and saving this code provides an audit trail for other psychometricians both to validate the data analysis you conducted and – if necessary – spot and debug any errors. In addition, the code is typically reusable for different exam administrations. This means that your item analyses don't have to be redesigned from scratch every time a new iteration of the exam is administered.

4 | Item and Test Analysis Reporting

Once the item and test analyses have been completed, their results need to be enshrined and disseminated to the examination stakeholders. This mainly involves running the 'clean' data through the appropriate data analysis programs; and obtaining a summary report on the statistics of interest (for example, the p-value for each item; the CRPB for each item).

Once these reports are generated, a psychometrician will normally review the results and make an interpretive summary of the findings (which will then be shared with the testing organization). The point of this summary is to provide a data-informed sense of the quality of the items (and the exam as a whole). A clear, well-informed summary at this stage is crucial in order to help guide the testing organization's decisions around retaining, revising, or retiring items from their exam.

In providing this summary, it is important to understand the difference between data-informed decision-making, and data-dictated decision-making. When reviewing any given item in an exam, the decision to retain, revise, or retire that item should be informed by the item-level statistics; however, in many cases, it may not be advisable to automatically decide to retire an item simply because it failed to meet a given p-value, CRPB, or other numerical cut-off value. Items that fail to meet these standards are generally flagged for review and discussion with the exam validation committee (or other members of the testing organization); however, the final decision about what to do with that item should be based on a holistic assessment of the item's content and adherence to the examination blueprint, in addition to the item-level statistics on candidate performance.

5 | Review and Interpretation of Analyses

Conducting a thorough and well-informed item analysis is something that requires training and experience. However, these processes are not magic. In fact, a few basic principles underlie the entire process.

In the next section, we'll try to demystify these processes; and provide some concrete, applied direction on how to understand and leverage the various analyses that go into conducting an item and test analysis.

TEST ANALYSIS

Test analysis involves reviewing the psychometric performance of an exam at a broad test level (that is, the general performance of the exam across all individual items). In doing so, a test analysis provides a "birds eye view" of the exam results; and ensures that everything is performing as expected in terms of the overall mean score that candidates received; the variation in overall exam scores; the reliability of candidates' exam scores; and the percentage of candidates who passed (i.e., met or exceeded the exam's cut score). These indices can often be effectively summarized in a single table like the one shown below:

EXAM STATISTICS

Reliability (KR-20)	0.96	**Average score**	76.32	**Standard deviation**	11.43	**Standard Error**	1.07

ABILITY GROUPINGS

Low	n = 31	**Medium**	n = 37	**High**	n = 32	**Total**	n = 100

Cut score = 65/100 (65%) **Pass rate = 85% (85/100 candidates)**

In this example, we can see an example of a fairly "good" set of test-level statistics. You'll notice there's a high KR-20 value of .96 (suggesting high reliability in exam scores); a mean exam score of 76.32, with a standard deviation of 11.43 (which are pretty typical for an exam with a cut score of 65.0%); and a pass rate of 85.0%.

Now, let's consider the results of a test analysis for an exam that that performed "poorly" (and would be much less likely to be see continued use in a certification/licensure context):

EXAM STATISTICS

Reliability (KR-20)	0.48	**Average score**	21.22	**Standard deviation**	3.76	**Standard Error**	1.87

ABILITY GROUPINGS

Low	n = 83	**Medium**	n = 39	**High**	n = 78	**Total**	n = 200

Cut score = 65/100 (65%) **Pass rate = 85% (85/100 candidates)**

Unlike the previous example, here the reliability of candidate scores is quite low (KR20 = .48). The average exam score is also quite low (21.22), as is the pass rate (2.0%). If we were to encounter these statistics during our test analysis, it would raise a few alarm bells – and may warrant a closer look at one's analysis code to ensure that the analysis was conducted properly. If it was conducted properly, these numbers suggest there are some serious problems with the examination's content (which should be brought to the testing organization's attention).

ITEM ANALYSIS: A CLASSICAL TEST THEORY APPROACH

Unlike test analysis (which is meant to provide a broad-level overview of an exam's performance), item analysis delves into the performance of specific items within the examination. In doing so, these analyses assess the extent to which each individual item in an exam is performing in accordance with the standards set forth by the examination policy group.

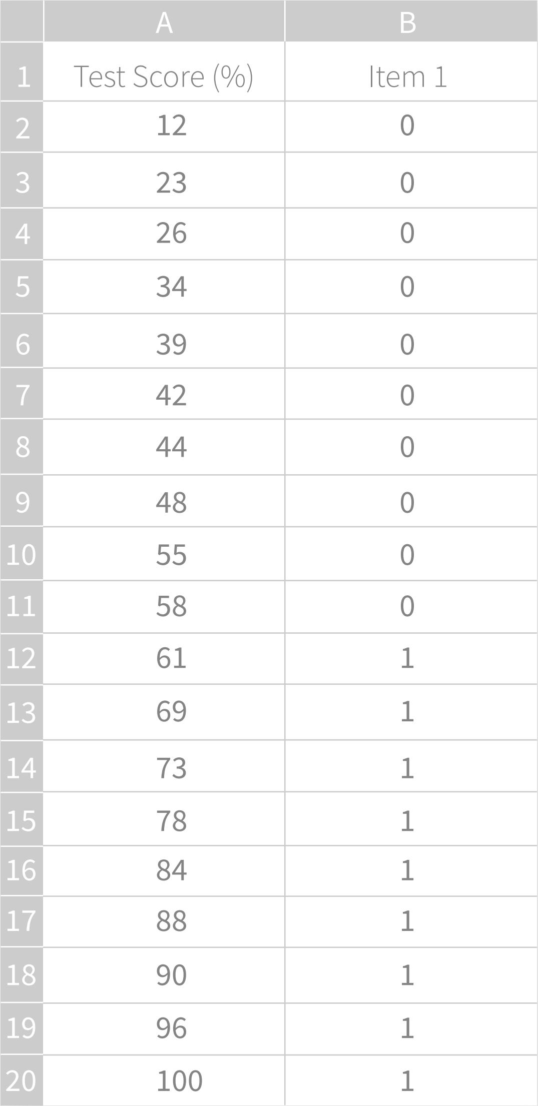
There are two main theoretical approaches to conducting an item analysis (or test analysis for that matter): Classical Test Theory (CTT) and Item Response Theory (IRT). Of these, Classical Test Theory is the more common and – in many cases – more popular approach, as it is easier to explain to stakeholders and doesn't require as large of a candidate pool as IRT does.

Although a comprehensive overview of CTT is beyond the scope of this paper, we will try to cover the essential information to look for when conducting a CTT-based item analysis. Normally, when conducting such an analysis, we consider the following key variables:

- Number of candidates: The more candidates that answer a question, the more robust or "stable" the statistical information for that question will be.
- p-value (i.e., item difficulty): This refers to the proportion of candidates who answered an item correctly. For example, a p-value of .85 signifies that 85.0% of candidates got the item correct (indicating a fairly easy question).
- CRPB (i.e., item discrimination): This refers to the correlation between candidates' scores on a given item, and their performance on the overall exam. A high corrected point-biserial correlation (e.g., CRPB = .35) indicates that candidates' score on that item and their overall test scores are related. This is something we would expect and want to see – otherwise, the item may be assessing something different than the rest of the exam.

- Proportion of low-, medium-, and high-ability candidates selecting each response option: By separating the candidate pool into different categories based on their total score (for example, bottom 1/3 of scores = “low-ability candidates”; middle 1/3 of scores = “medium-ability candidates”; top 1/3 of scores = “high-ability candidates”), we can look for different response patterns among different groups of candidates. For any given item, we would expect low-ability candidates to select the right answer less often than medium- or high-ability candidates do; and to select the distractors (i.e., wrong answers) more often. A different pattern than this may suggest that the item is phrased in a misleading way (tricking high-ability candidates), or is otherwise problematic.

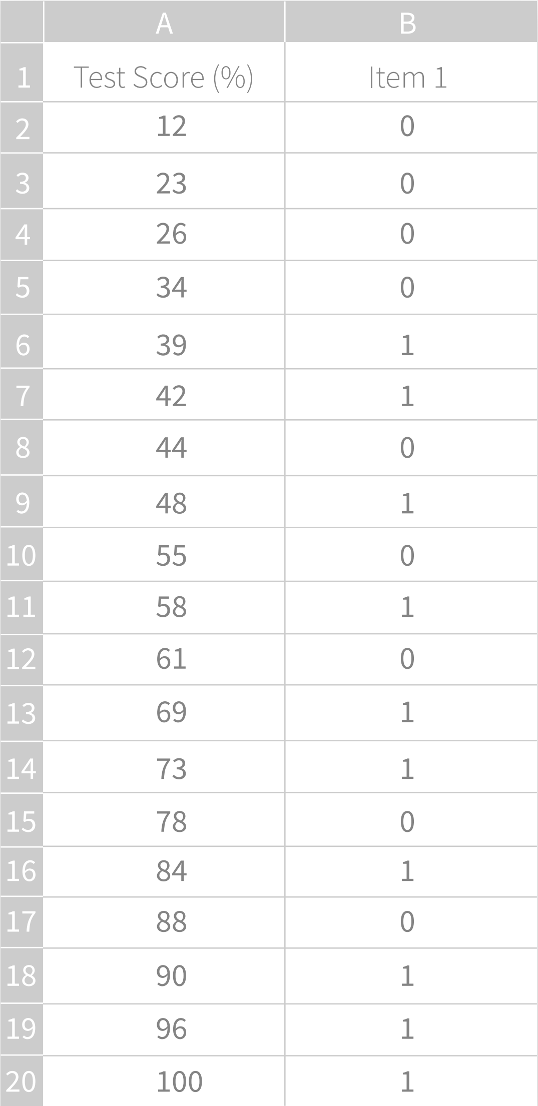
Of these, item discrimination is the primary indicator of item quality. The thinking behind discrimination is quite simple at its core: Namely, that you would expect candidates who get high test scores overall should be more likely to get any given item correct (relative to candidates who get low overall test scores). For example, consider the data shown:



	A	B
1	Test Score (%)	Item 1
2	12	0
3	23	0
4	26	0
5	34	0
6	39	0
7	42	0
8	44	0
9	48	0
10	55	0
11	58	0
12	61	1
13	69	1
14	73	1
15	78	1
16	84	1
17	88	1
18	90	1
19	96	1
20	100	1

The column on the left indicates each candidate’s total score on the exam; whereas the column on the right indicates whether or not that candidate got Item 1 correct or incorrect (0 = “incorrect”, 1 = “correct”). We can see that the candidates who got a “1” in the right column (indicating a correct answer to Item 1) also had higher scores on the exam overall. Indeed, the CRPB value for this example data would be .86 (extremely high!).

Of course, in the real world, item discrimination is not usually so clear-cut. Let's consider a more realistic example of an item with "good" discrimination (CRPB = .45):



	A	B
1	Test Score (%)	Item 1
2	12	0
3	23	0
4	26	0
5	34	0
6	39	1
7	42	1
8	44	0
9	48	1
10	55	0
11	58	1
12	61	0
13	69	1
14	73	1
15	78	0
16	84	1
17	88	0
18	90	1
19	96	1
20	100	1

So far, we've discussed item discrimination using the corrected point biserial correlation (i.e., CRPB). However, many data analysis programs will yield both a corrected (CRPB) and raw point biserial correlation (RPB). The only difference between these two indices is the way in which candidates' total exam score is calculated.

In the RPB index, each candidate's total score is calculated as the sum of their score across every item in the exam. However, there's a problem with this approach. Namely, this calculation includes the item that is being correlated with the candidate's total score (for example, the RPB of Item 12 in a 100-item exam will include candidates' score on Item 12 in its calculation of the candidate's total score). The correlation of a single item with itself is always perfect (i.e., 1.0). As a result, the correlation that is calculated between the item and the total score will be falsely inflated (that is, higher than it should be).

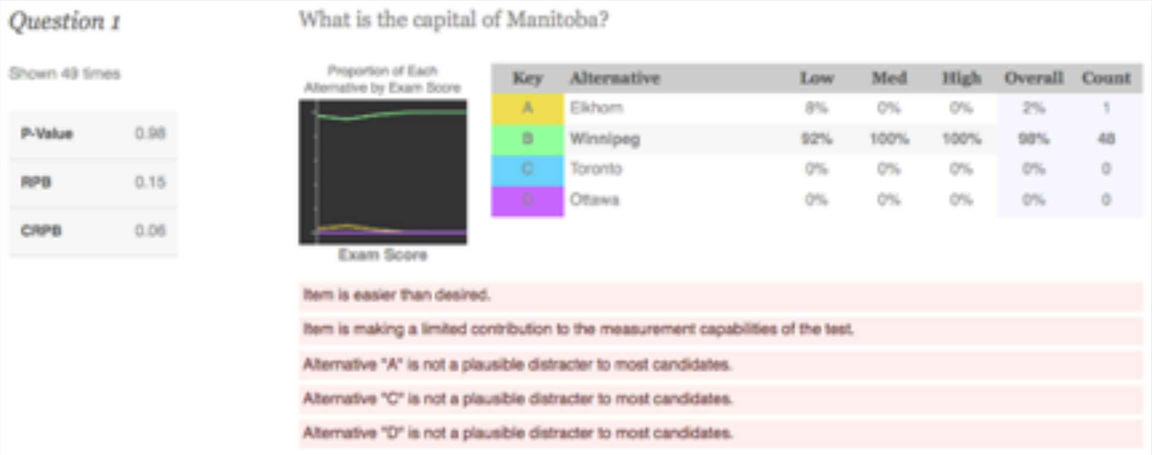
A CRPB accounts for this when calculating candidates' total exam scores. When calculating a CRPB, each candidate's total score is calculated as the sum of their score across every item in the exam except for the item being correlated with that total score. For example, in the same 100-item exam, the CRPB of Item 12 will include Items 1-11 and 13-100 in its calculation of each candidate's total score. As a result, the CRPB will always be smaller than the RPB. Because it makes this correction, the CRPB is generally considered to be more a valid and defensible index of item discrimination than the RPB.

SOME EXAMPLES OF CTT-BASED ITEM ANALYSIS

To help illustrate the concepts we've discussed so far, let's explore a few examples of item-level output; and see if we can diagnose the performance of the corresponding items using the CTT-based indices we've covered in this paper.

There are different ways to present the information subsumed under an item analysis. In these examples, we'll present high-level information about the item on the left; with more specific information about candidate response patterns on the right. There will also be a graphical representation of response patterns across various levels of ability (i.e., exam total scores).

Let's consider the first example below:



Question 1

What is the capital of Manitoba?

Shown 49 times

P-Value	0.98
RPB	0.15
CRPB	0.06

Key	Alternative	Low	Med	High	Overall	Count
A	Elkhorn	8%	0%	0%	2%	1
B	Winnipeg	92%	100%	100%	98%	48
C	Toronto	0%	0%	0%	0%	0
D	Ottawa	0%	0%	0%	0%	0

Item is easier than desired.

Item is making a limited contribution to the measurement capabilities of the test.

Alternative "A" is not a plausible distracter to most candidates.

Alternative "C" is not a plausible distracter to most candidates.

Alternative "D" is not a plausible distracter to most candidates.

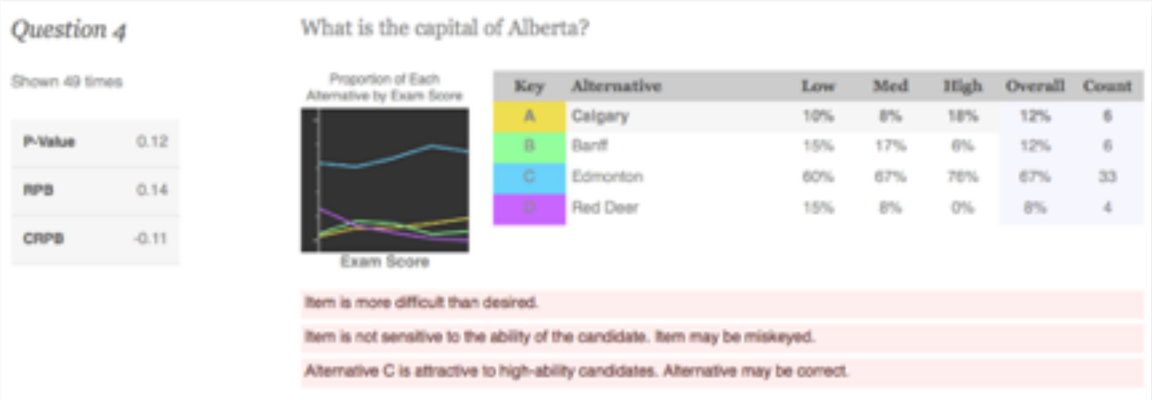
There are a number of issues to note with this item. For example, consider the following:

- Only 49 candidates responded to this question. This is a relatively small number of candidates. As a result, any item-level statistics will be less robust than desired. Having more candidates is always better for item analysis. Although there's no universal minimum number of candidates before we can "trust" the statistics, in general, having 100 or more candidates allows us to interpret these statistics with greater confidence.
- The p-value for this question is quite high (i.e., .98). This tells us that 98.0%of candidates answered this item correctly, suggesting the question it poses is very easy. Easy items like this can be a problem, because they don't help us discriminate between "competent" and "not-yet-competent" candidates; and therefore, offer a poor indication of what candidates know and can do.

- The discrimination indices (RPB and CRPB) are fairly low (CRPB = .06). This indicates that the item isn't doing a good job of discriminating between candidates who have high ability (i.e., those who did well on the test overall), versus candidates with low ability (i.e., candidates who did poorly on the exam overall). Low discrimination indices like this are fairly common when examining items that are either very easy or very difficult.
- The proportion of high-, medium-, and low-ability candidates that selected each answer option tells us that almost every candidate (regardless of their ability), selected the correct answer. Only one candidate in the low-ability category selected one of the distractors (Elkhorn). Based on this analysis, the distractor options do not appear to be very good at tempting low-ability candidates.

Overall, this item appears to be extremely easy, and unable to effective discriminate between high- and low-ability candidates. As a result, it would be of limited value on an exam.

Hopefully that helped you get a sense of how CTT-based item analysis works in practice! Let's consider another example below:



Question 4

What is the capital of Alberta?

Shown 49 times

P-Value	0.12
RPB	0.14
CRPB	-0.11

Key	Alternative	Low	Med	High	Overall	Count
A	Calgary	10%	8%	18%	12%	6
B	Banff	15%	17%	6%	12%	6
C	Edmonton	60%	67%	76%	67%	33
D	Red Deer	15%	8%	0%	8%	4

Item is more difficult than desired.

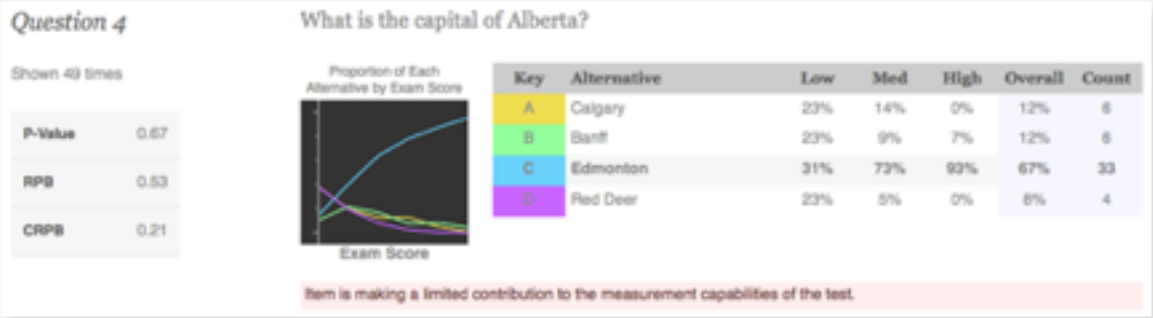
Item is not sensitive to the ability of the candidate. Item may be miskeyed.

Alternative C is attractive to high-ability candidates. Alternative may be correct.

In this case, we can see the following issues:

- Again, only 49 candidates responded to this question.
- The question appears to be very difficult (i.e., a p-value of .12; indicating that only 12% of candidates answered the question correctly).

- The discrimination indices (i.e., RPB and CRPB) are very low. In fact, the CRPB is actually negative (suggesting that lower-ability candidates who did poorly on the rest of the exam were more likely to get this item correct). A negative discrimination index is usually a major point of concern; and can indicate that the question may be mis-keyed.
- The proportion of high-, medium-, and low-ability candidates selecting each response option supports the idea that this item may be mis-keyed. Notice that Calgary has been keyed as the correct answer; however, the frequency with which this option was selected does not appear to differ much between the high-, medium-, and low-ability candidates. However, if we look at some of the other response options, Edmonton looks like it might be correct. When looking at the Edmonton option, we can see that low-ability candidates choose this option less often than medium- and high-ability candidates. Indeed, it turns out that Edmonton is the correct answer! All we need to do here is change the keyed correct answer to "Edmonton", and this item can be re-analyzed (see the corrected version below):


Question 4

Shown 49 times

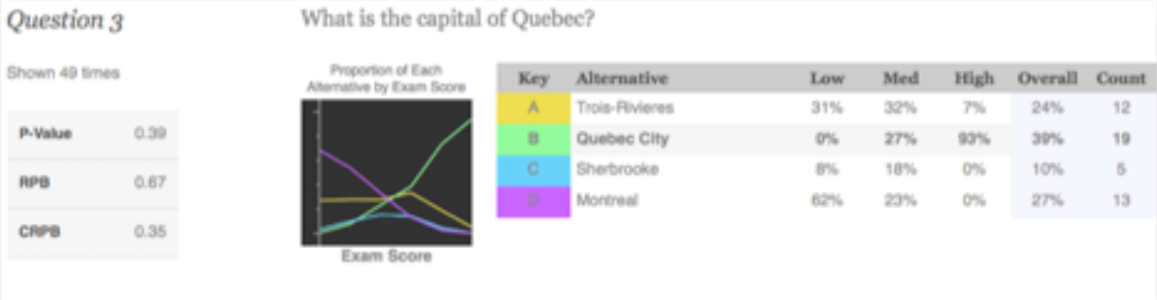
P-Value	0.67
RPB	0.53
CRPB	0.21

What is the capital of Alberta?

Key	Alternative	Low	Med	High	Overall	Count
A	Calgary	23%	14%	0%	12%	6
B	Banff	23%	9%	7%	12%	6
C	Edmonton	31%	73%	93%	67%	33
D	Red Deer	23%	5%	0%	8%	4

Item is making a limited contribution to the measurement capabilities of the test.

Excellent! Let's try our hand at another example (shown below):


Question 3

Shown 49 times

P-Value	0.39
RPB	0.67
CRPB	0.35

What is the capital of Quebec?

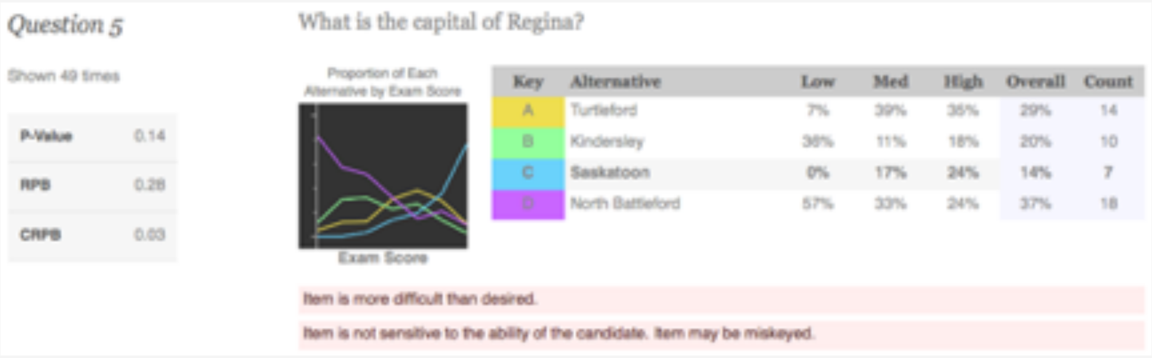
Key	Alternative	Low	Med	High	Overall	Count
A	Trois-Rivieres	31%	32%	7%	24%	12
B	Quebec City	0%	27%	93%	39%	19
C	Sherbrooke	8%	18%	0%	10%	5
D	Montreal	62%	23%	0%	27%	13

As you examine this item, you should notice the following:

- Only 49 candidates answered this question. Again, the candidate pool is rather small.
- The item appears to be rather difficult (i.e., a p-value of .39, meaning that 39% of candidates answered the question correctly).
- Although the question is difficult, the discrimination indices (i.e., RPB and CRPB) are fairly good (CRPB = . 35). This means that the item – though difficult – seems to do a good job of distinguishing between high-ability and low-ability candidates.
- The proportion of high-, medium-, and low-ability candidates who selected the correct response (Quebec City) looks good, overall. No low-ability candidates selected this option; whereas 27% of medium-ability candidates did (and 93% of high-ability candidates did). The distractor option that drew the most candidates was Montreal. This makes sense, as Montreal is a well-known city in Quebec. It seems that the low-ability candidates who didn't know the material just selected the city that they recognized; but this did not fool the high-ability candidates (the mark of a good distractor option!).

Moving forward, this item may benefit from a re-working of the distractors. Specifically, this item should be re-crafted by a subject matter expert to replace the two distractors "Kelowna" and "Salmon Arm" with more plausible (though still incorrect) options.

Let's consider one more example before we move on to IRT-based item analysis:



Question 5

Shown 49 times

P-Value	0.14
RPB	0.28
CRPB	0.03

What is the capital of Regina?

Key	Alternative	Low	Med	High	Overall	Count
A	Turtleford	7%	39%	35%	29%	14
B	Kindersley	38%	11%	18%	20%	10
C	Saskatoon	0%	17%	24%	14%	7
D	North Battleford	57%	33%	24%	37%	18

Item is more difficult than desired.

Item is not sensitive to the ability of the candidate. Item may be miskeyed.

Again, there are a few things to note about the performance of this item:

- Again, only 49 candidates answered the item.
- The item appears to be quite difficult (i.e., a p-value of .14, meaning only 14% of candidates answered the item correctly).
- The discrimination statistics (i.e., RPB and CRPB) are close to zero, suggesting that this may not be effective at differentiating between high- and low-ability candidates.
- An examination of the response patterns suggests that the candidates were quite confused. Candidates at all three levels of ability are all over the place in terms of which response option they selected (with no discernible pattern of responses based on ability). It seems the candidates have no idea how to answer this question!

And indeed, the question does not make sense. Regina is a city (not a province); so, asking candidates to identify the capital of Regina makes no sense. It is likely that the subject matter expert meant to ask: “What is the capital of Saskatchewan?” (and include “Regina” as an answer option). The content here would need to be re-crafted by a subject matter expert.

ITEM RESPONSE THEORY AND ITEM/TEST ANALYSIS

Although CTT-based approaches to item analysis are useful, relatively easy to understand, and quite popular, IRT-based analyses do afford some unique advantages.

For example, when taking a computerized adaptive test, candidates are first presented with an item of average difficulty. If the candidate answers the item correctly, they are given a more difficult item (whereas if they answer the item incorrectly, they are given a less difficult item). This process is repeated after each response in the exam, allowing the test to automatically “home in” on the candidate’s ability level. The difficulty indices used to categorize these items (and thus, enable the adaptive test to proceed), are often based on IRT; making this approach to item analysis invaluable to the development of adaptive testing environments!

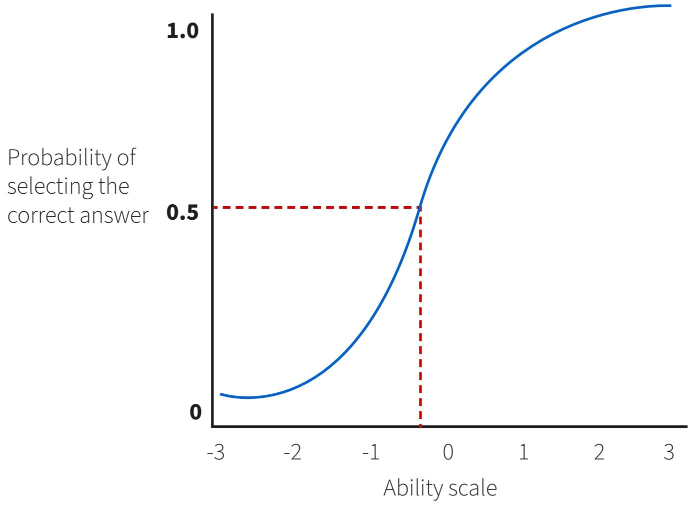
So how do these indices work (and how do they compare to the indices subsumed under CTT)? Although there are several different types of IRT analyses, they all centre around a formula and corresponding graph called an item characteristic curve (ICC). This graph visually depicts the probability of answering a given item correctly as a function of the candidate’s underlying ability (operationalized as theta, or ϴ).

In order to estimate this relationship, a number of different item-level parameters are modeled. Depending on the specificity of the question, sample size, and computing power available, there are more or less complicated models that a psychometrician can run. But one commonly-used IRT-based analysis is the '3PL' model (or three-parameter model). As its name would imply, this analysis estimates the relationship between a candidate's ability (i.e. ϴ) and the likelihood of that candidate correctly answering the item by modeling three item parameters:

- **a-parameter:** This represents the steepness of the "slope" on the item characteristic curve. For any given item, we see that, as ability increases, the candidate's likelihood of answering the item correctly also increases. However, different items will differ in terms of how sharply this increase occurs. Items with a steep slope (i.e., a higher a-parameter) indicate that, as ability increases, the likelihood of answering the item correctly sharply increases. Conversely, items with a gentle slope (i.e., a lower a-parameter) indicate that, as ability increases, the likelihood of answering the item correctly gradually increases. Thus, items with steeper slopes are better at differentiating candidates who have the requisite level of ability from candidates who do not (and thus, are more discriminating). Following this logic, you would ideally like to see a-parameters that are as high as possible.
- **b-parameter:** The b-parameter indicates the level of ability (i.e., ϴ) that a candidate would need in order to have a 50.0% likelihood of answering the item correctly. As such, this parameter corresponds to the item's difficulty. These values are reported as z-scores (which have a mean of zero). This means that any item with a b-parameter of more than 0.00 requires more than the mean level of ability to answer correctly; whereas any item with a b-parameter of less than 0.00 requires less than the mean level of ability to answer correctly. For example, an item with a b-parameter of -2.00 would be very easy; an item with a b-parameter of 0.00 would be of average difficulty; and an item with a b-parameter of 2.00 would be very difficult.
- **c-parameter:** Sometimes, even if the candidate has very low ability and doesn't know anything about the material, they can still guess the correct answer. The c-parameter is an index of guessing; and represents the probability of answering a given item correctly when the candidate has extremely low ability (or ϴ). On the ICC, this is represented as the point on the y-axis where the curve crosses the x-axis (i.e., the height of the line on the furthest point to the left of the graph). Higher points (or higher c-parameters) indicate that the item is more "guessable" than items with lower c-parameters.

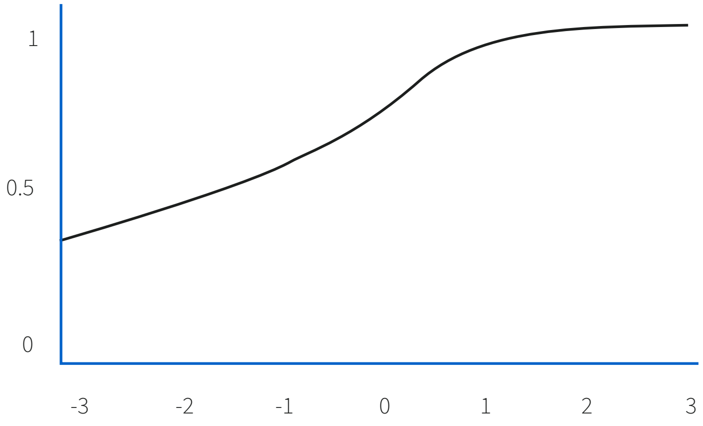
Here is an example of an ICC that was modelled using the 3PL approach. Note that it has the following parameters:

- **a-parameter:** 1.13 (High discrimination)
- **b-parameter:** -0.40 (Low difficulty)
- **c-parameter:** 0.02 (Low probability of guessing the correct answer)



Let's consider a few other examples (again, all modelled using the 3PL approach):

- **a-parameter:** 0.45 (modest discrimination)
- **b-parameter:** -0.70 (Low difficulty)
- **c-parameter:** 0.40 (High probability of guessing the correct answer)

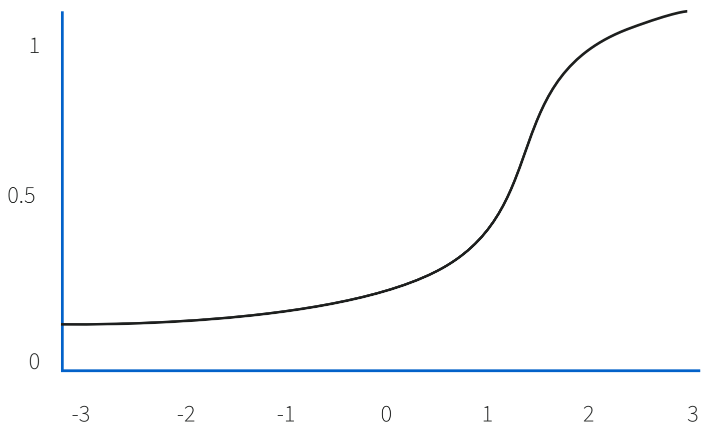


The ICC above gives us an example of a fairly mediocre item. We can see that the item has some ability to discriminate between high- and low-ability candidates (that is, the a-parameter is above zero, indicating the slope is positive); however, the slope is quite gentle, indicating that increases in ability only gradually lead to increases the probability of answering the item correctly.

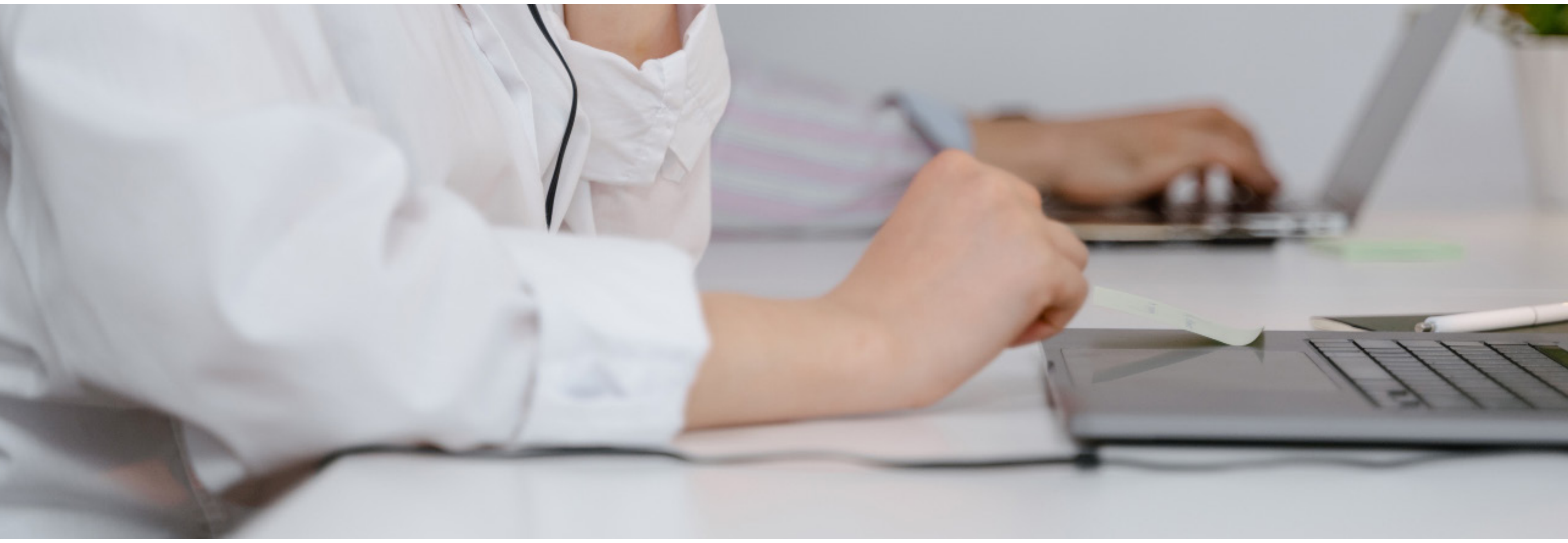
More concerning, however, is the fact that a large proportion of candidates with very low ability seem to be answering this item correctly (i.e., a c-parameter of .40). While it would be more desirable to have a steeper slope, the larger issue with this item would be understanding why the c-parameter is so high, and adjusting the item to make it less “guessable” (for example, it may require new distractors; or there may be clues in the stem that hint at the correct answer).

Hopefully that helped! Let’s try our hand at another example:

- **a-parameter:** 1.31 (Very discriminating)
- **b-parameter:** 1.49 (Very difficult)
- **c-parameter:** 0.18 (Low probability of guessing the correct answer)



This item would be considered high-quality (although also a little difficult, as evidenced by the b-parameter of 1.49). Difficulty notwithstanding, you’ll notice that the item still has a lot going for it. The a-parameter (or discrimination index) is very high; and only a small portion of candidates with very low ability managed to guess the correct answer.



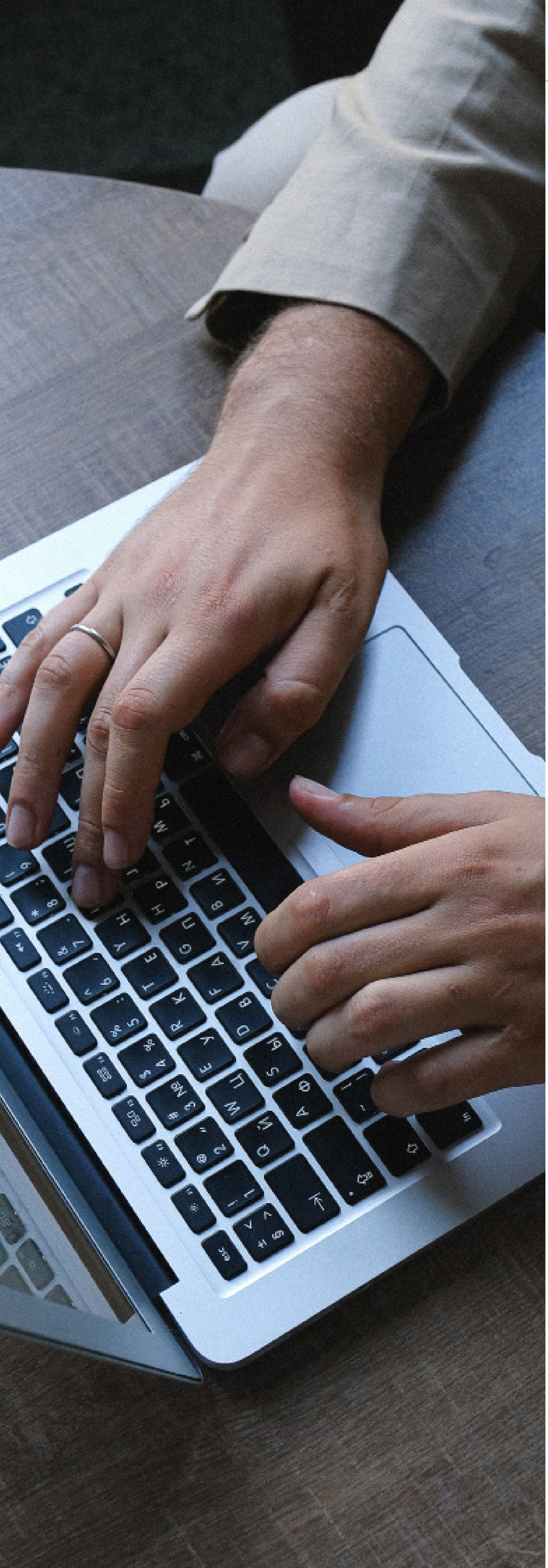
6 | Refining Item Development Targets

After the item and test analysis has been completed, it's time to survey the "damage" (so to speak). It's possible that, during the review, a number of poorly-performing items may have been retired from the item bank.

Based on these removal decisions, new item development targets may need to be set (in order to fill in the resulting gaps that were created during the review). Indeed, in some cases, a significant percentage of the questions for some blueprint categories may be rejected. Often this is due to the fact that some blueprint areas are just harder to write good questions for.

This is a normal part of the item development process. To help prepare for this eventuality, it's often a good idea to keep track of an exam's historical item rejection rates. In doing so, you can better anticipate and plan for ongoing item development efforts. For example, if for a specific examination program, the historical rejection rate has been 30%, you might plan to set initial item development targets at least 30% higher than would be required to fulfill the exam blueprint (as this accounts for the number of items required to fill gaps in the blueprint as well as the likely number of items that will not survive the item analysis process).

It should also be mentioned that even the best questions have a 'shelf life'. For content security reasons, a question should only appear a maximum number of times (or be seen by a maximum number of candidates on an operational test form), before it is retired from the item bank. In this sense, the item development process is iterative in nature; and never really ends, as long as the assessment program is still running.



Next Stage | Scoring and Reporting

Now that your candidates have finished writing the exam and you have a final list of items to be included or excluded from scoring, it's time to compile the final performance reports!

The next stage of the Assessment Life Cycle involves running the "clean" candidate data through a data analysis program to generate final exam scores; and then generating various candidate performance and summary reports to share with the candidates, testing program, and schools the candidates came from. This stage will be discussed in more detail in the next white paper in this series.

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 fifth of these stages (i.e., item and test analysis). As covered, this involves a number of specific steps, including:

- Establishing standards for item quality and item reliability
- Data export and cleaning
- Code preparation
- Item and test analysis reporting
- Review and interpretation of analyses
- Refining item development targets

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 5

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 5: Item and Test Analysis:**

Service	Description
Classical Test Theory (CTT) item analysis	Classical Test Theory (CTT) is by far the most commonly-employed approach to item and test analysis. CTT provides useful diagnostic information regarding the performance of items and – more broadly – exams (even when dealing with a limited number of candidates). At Meazure Learning, we offer tailored CTT-based analyses to our clients, which we can use to effectively diagnose the psychometric performance of exam items. Accurately interpreting and reporting on the results of these analyses is crucial to ensuring the defensibility of an examination program; and can provide invaluable feedback regarding areas where item development can be improved.
Item Response Theory (IRT) item analysis	Item Response Theory (IRT) provides rich information regarding items, and is highly useful for a wide range of assessment situations (e.g., computer adaptive testing). Clients with access to larger candidate pools are in an excellent position to leverage the potential of IRT-based item and test analysis to help evaluate the performance of their item bank. At Meazure Learning, we offer a wide range of in-house expertise to help our clients leverage the potential of IRT analytics. This includes specific services for using IRT in adaptive testing environments, differential item functioning analysis, and collusion detection.

Differential Item Functioning (DIF) analysis

Differential Item Functioning (DIF) analyses allows a testing program to assess how the items in their examination perform across different groups of candidates (for example, if both English- and French-speaking candidates answer questions in the same way).

Ensuring that your items are fair for all groups of candidates is essential to ensuring the validity and defensibility of your assessment program. Let Meazure Learning help you ensure this defensibility through the wide range of DIF analyses and expertise we offer to our clients.