Another Approach in Item Response Theory
The “New” Kid on the Block

What you’ll learn in this mini-chapter:

- A brief history of item response theory
- What item response theory is and how it is used
- Simple item response theory output and what it means

Item response theory is not exactly new, but it is just about the newest popular approach to the design, administration, and evaluation of tests. Researchers from the Biometrics Unit of the Institut Régional du Cancer Montpellier in France compared item response theory with the more classical methods of testing to analyze health-related quality-of-life data. The researchers reached the same conclusions with both tests; no difference was found between the two different treatments. They concluded that the item response theory model may be more complex but shows the same qualities and gives similar results. It has the advantage of being more precise and suitable because of its direct use of raw data. Pretty neat.

THE BEGINNINGS OF ITEM RESPONSE THEORY

In the first five chapters of this book, you learned quite a bit about the underpinnings of some of the basic concepts upon which much of classical test theory is based. Notice the word classical in the previous sentence, the study of which started sometime around 1920. And for the past almost 100 years, these approaches have been the basis for test development, administration, and evaluation.

But things sure do change in every discipline, and so it is in the field of tests and measurement.

This brief chapter is about item response theory (IRT). IRT is a particular perspective on how test items should be developed and evaluated as they are refined to best take advantage of an individual's “true” ability. So when discussing IRT, rather than focus on the individual (and sources of error such as method error or trait error, as we discussed in Chapter 3), we focus on the individual test items and how well they discriminate between those test takers who (in theory) know the material and those who do not.

What's so cool about IRT, and why is it such an attractive alternative to classical test theory? Most uniquely, it focuses on and estimates the ability of the test taker independent of the difficulty of the items. IRT does this by looking at the relationship between the performance on each individual item on the test and the underlying ability of the test taker. Why underlying? Because the level of ability is not explicit; it's never a known quantity.

IRT Theory, 1; Classic Test Theory, 0. Let's say there are 11 different possible scores (0 correct, 1 correct, 2 correct, etc., up through 10 correct—so 11 possibilities). But in IRT, there are as many possible outcomes as there are possible item combinations, so one possible outcome could be Item 1 correct and the rest of the items incorrect. Or Items 1 and 2 correct and the rest incorrect.
In fact, the possible number of combos is the possible outcomes raised to the power of the number of items—in this example, $2^{10}$ or 1,024. So rather than a mere 11 data points for performance, there are now 1,024—a much finer scale on which to grade outcomes, allowing more precision and better decisions.

The Power of IRT. The development of IRT was based mostly on the work of Robert Lord, Melvin Novak, and Alan Birnbaum in their seminal papers, first published in the ’60s and ’70s and then brought to the attention of the tests and measurement world. So if IRT is so cool and useful, why did it take so long to become popular and why isn’t it even more popular now? Easy—a simple one-word answer: computers. The use of IRT requires huge amounts of data, and the more data, the more effective and accurate the analysis. There also need to be at least 500 individual test results—and thousands more are commonplace—for the results of the analysis to be trustworthy. This size requirement is just not practical for the average classroom- or workplace-designed test. But don’t be surprised if you see thousands of test takers per item in some IRT analysis. The more of such, the more precise and reliable the outcomes.

But what about this item–ability relationship?
We start with the assumption that everyone has some level of ability on a specific topic when being tested. Whether you score 100% or 50% correct on an achievement test, this score acts as an indirect measure of ability. And while it may not reflect the true ability of the individual, as we know from classical test theory, this score is a starting point.

This underlying ability, which is critical to the theory, is termed a latent trait (meaning hidden or unobservable), and while present, it is not obvious. But even with an observed score of 100% or 50%, test scores are only estimates and more or less correspond to the true score or ability of the individual. But as you know from Chapter 3, this obtained score is only an estimate, and true score remains a theoretical concept. The job of the modern psychometrician is to estimate as accurately as possible one’s true underlying latent trait score. Many scientists believe that IRT can do a better job at this than the techniques associated with classical test theory.

In effect, IRT considers each individual item, not just the total score, and looks at the patterns of responses across all items and all test takers.
The most fundamental aspect of understanding IRT is what is called the item characteristic curve (ICC), shown in Figure 6.1.

Let’s take a look at this sample curve and what it represents. Remember, this is the curve for just one item on a test, and for each item there is one item characteristic curve.

The *x*- or horizontal axis represents the latent or underlying trait or ability that the individual test taker brings to the item itself, and the “amount” of trait is expressed along that axis. This underlying ability is called theta, represented by the Greek symbol \( \theta \). Average ability is located at 0 on the \( x \)-axis (that’s the mean ability level), average ability to the right, and below-average ability to the left.

The *y*- or vertical axis is the probability of getting the item correct—and that’s represented as \( P(\theta) \)—or the probability of a correct response given a certain level of ability or \( \theta \). You can see it ranges from 0 to 1, or 0% to 100%. The lower this value, the more difficult the particular item. The higher the value, the more likely that the test taker will get it correct.

Let’s put \( x \) and \( y \) together.

If you follow the curve, you can see how the more ability an individual brings to a particular item, the more likely (the higher the probability) that he or she will see success on this one item. The higher the value of \( \theta \), the more likely that an individual will get the item correct.

At the lowest levels of ability, the probability of getting a correct response should be very low. At the highest levels of ability, the probability
of getting a correct response should be very high. That should be apparent through an examination of the curve in Figure 6.1. And again, this is the curve for only one hypothetical item.

**TEST ITEMS WE LIKE—AND TEST ITEMS WE DON’T**

Within IRT, there are two characteristics of any one item that distinguish items from one another and also allow us to pass judgment on whether the item is a good one. Remember that in the IRT world, good items discriminate differently at different levels of ability—between those who know the material on the test and those who do not.

The first characteristic is the difficulty level of an item as noted by the location of the curve along the x-axis. The farther to the right the curve lies (toward higher ability), the more difficult the item. The farther to the left (toward lower ability), the easier the item. For example, in Figure 6.2, you can see how the probability of getting the item correct increases as theta increases; Item 1 is easier than Item 2, which in turn is easier than Item 3. It takes less ability to get Item 1 correct than it does to get Item 2 correct than it does to get Item 3 correct. In other words, given the same level of ability ($\theta = .5$, in this example), the probability of getting the item correct varies as a function of difficulty level.

![Figure 6.2 How Items Differ in Their Difficulty Level](image-url)
The discrimination level of an item, the second characteristic, is reflected in the steepness of the item characteristic curve. The steeper the curve, the better the item discriminates or distinguishes between those test takers who know the material sampled by the one item and those who do not. As you can see in Figure 6.3, the steeper the curve, the better the item discriminates; the flatter the curve, the poorer the item discriminates.

The curve for Item 1 is steep and discriminates well. You can see that for the same degree of ability ($\theta$), the probability of success can vary greatly. Likewise, Item 2 does not discriminate well, since the probability of a correct response is very similar regardless of underlying ability.

What would perfect discrimination look like? Take a look at Figure 6.4. Here, the ability level is the same ($\theta = 2.0$) and the item discriminates perfectly at that ability level. That is, all the individuals with a theta less than 2 (lower-ability folks) have a near 0% probability of getting the item correct, and all the individuals with a theta more than 2 (higher-ability folks) have a 100% chance of getting the item correct.
To better understand the characteristics of an IRT curve, three different characteristics have been defined, and each curve can differ from others on one or all of them.

The slope of the curve, which we mentioned earlier, is the discrimination level, defined by $a$. When the curve is steeper, there is a large difference in the probability of a correct response for those whose theta values differ. When the curve is flatter, there is little distinction.

The point along the x-axis that defines the difficulty level of the items is $b$. A higher value of $b$ indicates that the item is more difficult than a lower value of $b$. When the difficulty level is less than 0, the likelihood of a correct response is greater than .5 (these are probably easy items), and when $b$ is greater than 0, the likelihood of a correct response is less than .5 (these are probably hard items).

The probability of a correct response for test takers who have low ability and of getting the item correct when guessing is $c$.

Lawrence Rudner presents a terrific applet for varying these values and seeing what the item characteristic curve looks like, and you can find it at https://archive.is/5yxWF. We can use it to look at some of the following examples of how the item characteristic curve changes. For our purposes and at this basic level, we are going to pay attention to only the $a$ and $b$ values.
In the following item, discrimination or $a$ is 2.0, and difficulty or $b$ is –2.5.

And this item has an $a$ value of 2.0 and a $b$ value of 2.0 as well:

We can use the values of $a$ and $b$ to make decisions about the worthiness of an item. And as those items are used and evaluated, they can be refined, placed once again in the test pool of items, and reevaluated until they meet the criteria we use to define good items.

For example, the first item above is an “easy” item with a low level of difficulty equal to –2.5 (you don’t need to know much to have a high probability of getting this item correct) and does not discriminate very well between test takers. On the other hand, the second item is a more “difficult” item with a difficulty level equal to 2.0 (see how the curve is placed toward the right side of the $x$-axis?) and discriminates well, with a pronounced slope.
Putting $a$, $b$, and $c$ Together

The values of $a$ and $b$ (and $c$) are mostly used in the assessment of how well items are working and which need to be revised.

The steps for the creation of a test consisting of many items and using IRT as a tool would be as follows:

1. Items are created by the test developer.
2. Items are included on a test.
3. The test is administered.
4. IRT is used to evaluate the usefulness of each item, one by one, and how well it “works.”
5. If necessary (which it almost always is), the test item is refined, rewritten, redrafted, etc., until there is another opportunity to use the item once again and revise as necessary according to the new values of $a$, $b$, and $c$.

When is this process of item (and test) development complete? When each item fits the difficulty and discrimination level that the test author feels adequate to discriminate between those who know the material and those who don’t.

There is another ICC function that assesses the amount of information revealed through an IRT analysis. It’s called the information function but is beyond the scope of our discussion here.

**IRT and Computerized Adaptive Testing.** Perhaps the neatest thing about IRT is its role in computerized testing or computerized adaptive testing. Knowing the difficulty and discrimination levels of items (after they have been developed and tested) allows the test administrator to adjust what subsequent items an individual will have to answer, based on IRT analysis of these items when given to thousands of previous participants. If an item is too difficult (Larry gets it wrong), the computer program can adjust which item is presented next—perhaps one that is easier. The computer program will adjust the presentation of items based on difficulty and discrimination levels to maximize the assessment of the individual’s underlying or latent ability. This means that everyone, in effect, will take a different test (or at least respond to a different set of items).
ANALYZING TEST DATA USING IRTPRO

As we mentioned earlier, computers were and are critical to the use of IRT. The analysis simply cannot be done without the help of such devices.

And as computers have become smaller, more powerful, and more affordable, the software to conduct very sophisticated analyses has become more accessible as well. Such is the case with IRT, and while there are many different programs available to conduct IRT analyses, we chose to show you how such an analysis can be done with IRTPRO (you can download a trial version from Scientific Software International at http://www.ssicentral.com/irt/). IRTPRO is available for Windows 10—sorry, no Mac version.

The sample we used for analysis is the result of a 20-item test taken by 500 individuals. Each item can be scored as right (1) or wrong (0). You can see a sample of the item scores in Figure 6.5, where columns are items, rows are individuals, and cell entries are correct (1) or incorrect (0) for any one item. For example, for Individual Test Taker 12, Item 3 was correct (there is a value of 1 entered in that cell).

In Figure 6.6, you can see the Unidimensional Analysis dialog box, where the items were added to the analysis. Then the RUN button was clicked, and in Figure 6.7, you can see the results of the analysis. Sophisticated software programs such as IRTPRO produce a large amount of output, but for our purposes, the output shown in Figure 6.7 gives us lots of information about which items we might keep and which really need to be revised.
Figure 6.6  The Unidimensional Analysis Dialog Box

Figure 6.7  Some IRTPRO Output

| Figure 6.7 Some IRTPRO Output | IRTPRO Version 3.0  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Output generated by IRTPRO estimation engine Version 5.10 (32-bit)</td>
<td></td>
</tr>
<tr>
<td><strong>Project:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Description:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Date:</strong></td>
<td>31 July 2016</td>
</tr>
<tr>
<td><strong>Time:</strong></td>
<td>05:18 AM</td>
</tr>
<tr>
<td><strong>Table of Contents:</strong></td>
<td></td>
</tr>
<tr>
<td>2PL Model Item Parameter Estimates for Group 1, logit: (a\theta + c) or (a(\theta - b))</td>
<td></td>
</tr>
<tr>
<td>Summed-Score Based Item Diagnostic Tables and (X^2)'s for Group 1</td>
<td></td>
</tr>
<tr>
<td>Group Parameter Estimates</td>
<td></td>
</tr>
<tr>
<td>Marginal (X^2) and Standardized LD (X^2) Statistics for Group 1</td>
<td></td>
</tr>
<tr>
<td>Item Information Function Values for Group 1 at 15 Values of (\theta) from -2.8 to 2.8</td>
<td></td>
</tr>
<tr>
<td>Likelihood-based Values and Goodness of Fit Statistics</td>
<td></td>
</tr>
<tr>
<td>Summary of the Data and Control Parameters</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>2PL Model Item Parameter Estimates for Group 1, logit: (a\theta + c) or (a(\theta - b)) (Back to TOC)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item</strong></td>
<td><strong>Label</strong></td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>Item 1</td>
</tr>
<tr>
<td>2</td>
<td>Item 2</td>
</tr>
<tr>
<td>3</td>
<td>Item 3</td>
</tr>
<tr>
<td>4</td>
<td>Item 4</td>
</tr>
<tr>
<td>5</td>
<td>Item 5</td>
</tr>
</tbody>
</table>
For example, in Figure 6.7, Item 1 is shown to discriminate well \((a = .72)\) and is of about average difficulty \((b = -.08)\). On the other hand, Item 4 does not discriminate well \((a = -.05)\) and is very easy \((b = -2.99)\). As the test developer, it would be your decision which items you want to keep in the item bank and, of those, which should be revised before they are used again.

**Seeing Is Believing**

It's enough to know what the values of \(a\), \(b\) and \(c\) are, but it's even better to be able to see them as a representation of how “good” an item is.

In Figure 6.8, you can see the actual IRTPRO output for an ICC of what is a relatively decent item. Much like the “perfect” ICC curve we showed you in Figure 6.1, the curve for this item seems to fit many of those characteristics, as well as discrimination and difficulty level. In contrast, in Figure 6.9 you can see the graphical representation of an unattractive item, one that has little discrimination power and difficulty that does not change as a function of ability.
SUMMARY

Okay, this is not the easiest material to understand, and this has been a very brief overview of what item response analysis is and how it might be used in tests and measurement practice. IRT’s main advantage is that it examines item responses instead of people responses; the emphasis is on the characteristics of the items and not the sources of errors emanating from people taking the test or the physical structure of the test itself. Even though you may not use it, you need to know about it, and this brief introduction (hopefully) informed you of the basics.

TIME TO PRACTICE

1. What is IRT, and what advantages does it have over classical test theory?

2. Why does IRT need a large number of respondents to be a valid way of analyzing test responses?
3. Go to the library (in person or online) and find an article that used IRT in the analysis. Then answer the following questions:
   a. What was the purpose of the study?
   b. How was IRT used?
   c. What conclusions did the author(s) reach given the IRT analysis?

WANT TO KNOW MORE?

Further Readings

IRT is primarily used in the measurement of ability and achievement and not as much in the area of clinical assessment. This article reviews the use of IRT in those clinical areas and the benefits, and discusses how the use of IRT in clinical assessment settings may hold great promise.


Here’s another introduction to IRT that appears in a journal not targeted at psychometricians or tests and measurement experts but, rather, people who focus on adolescent development. Often, these reviews targeted at audiences other than the primary researchers in the field can be very helpful.

And on the Internet
- Tips on IRT and how it can be applied and when it is best used can be found at https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf, in a paper titled “Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It.”
- Everything is easier to understand when you can “see” it. A very interesting visual guide to item response theory by Ivailo Partchev can be found at http://www2.uni-jena.de/svw/metheval/irt/VisualIRT.pdf.

And in the Real Testing World
One of the most interesting things about the development and dissemination of new techniques and ideas is how these techniques spread to different disciplines. So while IRT was initially the provenance of test makers and evaluators, its use has become commonplace in a variety of disciplines. In this study, the author talks about the use of IRT within the framework of studying adolescence.