Research Memorandum
ETS RM–20-06

Basic Concepts of Item Response Theory: A Nonmathematical Introduction

Samuel A. Livingston

June 2020

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Basic Concepts of Item Response Theory: A Nonmathematical Introduction

Samuel A. Livingston
Educational Testing Service, Princeton, New Jersey

June 2020

Corresponding author: S. A. Livingston, E-mail: slivingston@ets.org

Suggested citation: Livingston, S. A. (2020). *Basic concepts of item response theory: A nonmathematical introduction* (Research Memorandum No. RM-20-06). Educational Testing Service.

Action Editor: John Mazzeo

Reviewer: Tim Davey

Abstract

This booklet is a conceptual introduction to item response theory (IRT), which many large-scale testing programs use for constructing and scoring their tests. Although IRT is essentially mathematical, the approach here is nonmathematical, in order to serve as an introduction on the topic for people who want to understand why IRT is used and what it does but are uncomfortable with the mathematics. The focus is on the basic types of IRT that are commonly used today.

Keywords: item response theory (IRT), models, calibration, number-correct scoring, pattern scoring, ability estimates, score scales

Why This Booklet?

In the world of educational testing, an item is a question, problem, or task presented to a test taker as part of the test. Many large-scale testing programs use item response theory (IRT) for constructing and scoring their tests. IRT is essentially mathematical, and the books (and other materials) for teaching it take a mathematical approach. However, there may be people who want to understand why IRT is used and what it does but are uncomfortable with the mathematics. The purpose of this booklet is to offer these people an introduction to the basic concepts of IRT, without the mathematics. There will be no formulas or equations in this booklet. The explanations will consist of words and graphs. The language will be conversational, not academic.

I also hope that this booklet will be useful as a conceptual introduction to IRT for readers who do want to go on to learn the mathematics of IRT. If, after reading this booklet, you want a deeper understanding of IRT (including the math), there are several sources you can turn to.

This booklet is about IRT as it is now widely used, in the year 2020. For the past few years, some IRT experts have been working to create more advanced versions of IRT. Those more advanced versions are more complex than the versions of IRT that are in common use today. They also require much more data. This booklet will focus on the basic types of IRT that are commonly used today.

What Is IRT Good for?

IRT has several uses in making and scoring tests. It is useful for selecting items for a new form (edition) of a test and for pretesting new items to be used in future test forms. It is useful for producing scaled scores—scores that take into account the differences in the difficulty of the items on different forms of the test. These tasks can be done without IRT, but using IRT makes the job easier. It gives the test developers greater flexibility in choosing items for a new form of a test.

IRT can be used to determine the scaled scores on a new form of a test without waiting for data from people taking that new form. If the test is administered by computer, IRT can make it possible to report a scaled score for each test taker immediately after the test taker finishes the test. IRT also can make it possible to have the computer custom build a test form for each individual test taker. There can be 50 or more people taking the test at the same time, each one receiving a different set of randomly selected items. IRT makes it possible to report a score for each person in a way that accounts for the difficulty of the items on that person's test.

What Is IRT?

IRT is a mathematical theory about what happens when people take tests. It is all about probability—the probability that a test taker responding to a test item will answer it correctly.

What is probability? You may be surprised to learn that statisticians do not all agree on the answer to this basic question. For the purpose of this booklet, you can interpret probability as a number that indicates the chance that something will happen. That number represents a percentage of all the situations in which certain conditions exist. When the weather forecast says that the probability of rain today is 70%, the forecasters are saying that it will rain on 70% of the days with weather conditions like those today. When IRT says that a test taker's probability of answering a particular item correctly is .70, it is saying that 70% of the test takers like this one would answer this item correctly. But what, exactly, does "test takers like this one" mean?

The Main Assumption of IRT

Now we come to the main assumption of IRT. The assumption is that the probability that a test taker will answer a test item correctly depends on only one characteristic of the test taker. That characteristic is called "ability," which is a short way of saying "the knowledge or skill (or other quality) that this test measures."[1]

This assumption is often called "unidimensionality." It means that the test measures a single dimension. It means that if we knew a test taker's ability, that single number would tell us all that we could possibly know about the test taker's chance of responding correctly to any item on the test.

Of course, many real tests that use IRT do not actually meet this assumption. For example, a history test might have several items on government, several items on technology, several items on arts and culture, and so on. Such a test would not be truly unidimensional, but it could be close enough for IRT to be used effectively.

IRT Models

Most books and articles about IRT use the word "model." In ordinary language, a model is a representation of something—possibly a historic railroad train or a proposed building or a new type of equipment. The model has some selected qualities of the thing it represents—the qualities that are important for the model to do what it is intended to do. Often it is enough for the model to look like the thing it represents, but some kinds of models have to perform like the things they represent.

In statistics, a model is a representation of something that happens in the real world. The model consists of a set of mathematical statements. Those statements are usually assumptions about the relationships between things that can be measured. Typically, adding more statements to the model will provide a better description of the process that is being modeled. The model never provides a complete, exact description of the real process, but it can provide a lot of information about it. An eminent statistician once said, "All models are wrong, but some are useful" (Box, 1979).

The simplest version of IRT is the "one-parameter model" (also called the "Rasch model"). It assumes that the probability of a test taker answering an item correctly depends on only one characteristic of the item—its difficulty. Of course, that assumption often is not true. Nevertheless, this version of IRT is used on many tests, and it works fairly well most of the time.

Another commonly used version of IRT takes account of two characteristics of the item—its difficulty and its discrimination. Discrimination is the extent to which the item separates test takers above some point on the ability scale from test takers below that point. What is that point? That depends on the difficulty of the item. The more difficult the item, the higher the ability value where it discriminates most effectively. The version of IRT that uses these two characteristics of the item is called the "two-parameter model." It requires more data than the one-parameter model, but it often does a better job of producing scores that reflect each test taker's ability.

There is also a "three-parameter model" that uses three characteristics of the item—its difficulty, its discrimination, and how easy it is for a low-ability test taker to guess the correct answer. This version of IRT requires even more data than the two-parameter model, and its results tend to be very similar to those of the two-parameter model.

All three of these IRT models assume that there is a particular mathematical formula for the probability that the test taker will answer a test item correctly. The input variables for this formula are the test taker's ability and one, two, or three characteristics of the item:

- one-parameter model: item difficulty only,

- two-parameter model: item difficulty and discrimination, and
- three-parameter model: item difficulty, discrimination, and ease of guessing.

These IRT models assume that nothing else about the item has any effect on the probability that the test taker will answer the item correctly.

Nothing else? What about the position of the item in the test—near the beginning, in the middle, or near the end? What about the other items presented immediately before this item? We know that these factors can make an item easier or more difficult, and often they do. When we use IRT, we are assuming that those effects are small enough to disregard. One way to make sure they are small enough to disregard is to make sure they never change—at least, not enough to matter. The position of an item in the test won't matter if that item is always in approximately the same position (e.g., near the beginning of the test).

The mathematical formula that these models use is called a "logistic" function. A graph of it looks like Figure 1. The horizontal axis represents the test taker's ability. The numbers on the ability scale are usually chosen to represent some population of test takers, with 0 (zero) representing the average ability of the population. The height of the curve indicates the probability that the test taker will answer the item correctly. The height of the curve is close to zero for a very low-ability test taker. As the test taker's ability increases, the curve rises. It rises gradually at first, then it rises faster, and then it gradually levels off as the probability approaches 1.00. (A probability of 1.00 represents 100% certainty that the test taker will answer correctly.) This kind of graph, with an ability measure on the horizontal axis and the probability of a correct answer on the vertical axis, is called an "item response curve."

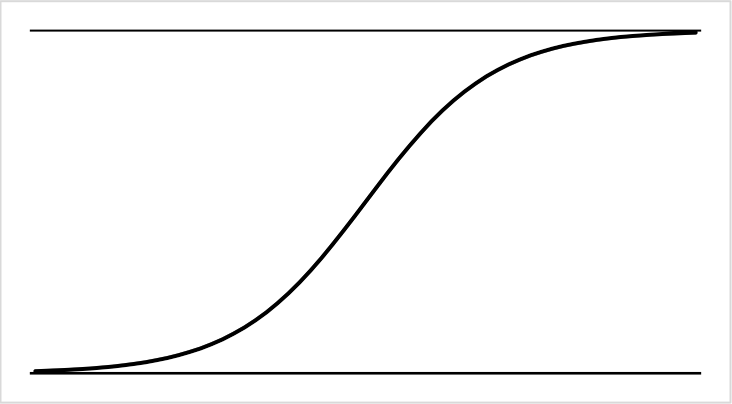

Figure 1. An item response curve.

The one-parameter model assumes that items on a test can differ in only one way that affects any test taker's probability of answering correctly: difficulty. Figure 2 shows the one-parameter item response curves for an easy item, a medium-difficulty item, and a difficult item. You can see that the curves all look the same, except for their position on the ability scale. In the one-parameter model, if we know the position of the curve on the ability scale, we have all the information we need to determine the entire curve.

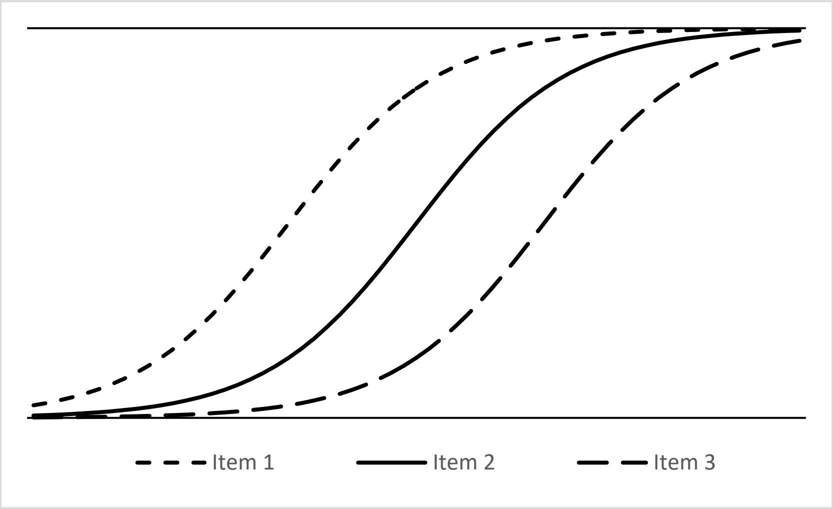

Figure 2. Item response curves in the one-parameter model.

The two-parameter model assumes that items can differ in difficulty and in discrimination (but not in any other way that affects any test taker's probability of answering the item correctly). Figure 3 shows the two-parameter response curves for two items. Notice that Item 1 is much easier than Item 2 for low-ability and middle-ability test takers—it gives them a much higher probability of a correct answer. But both items are easy for high-ability test takers. Also notice that Item 1 discriminates gradually over a wide range of ability. Item 2 discriminates more sharply but in a narrower range of ability. Outside that narrower ability range, it does not discriminate at all.

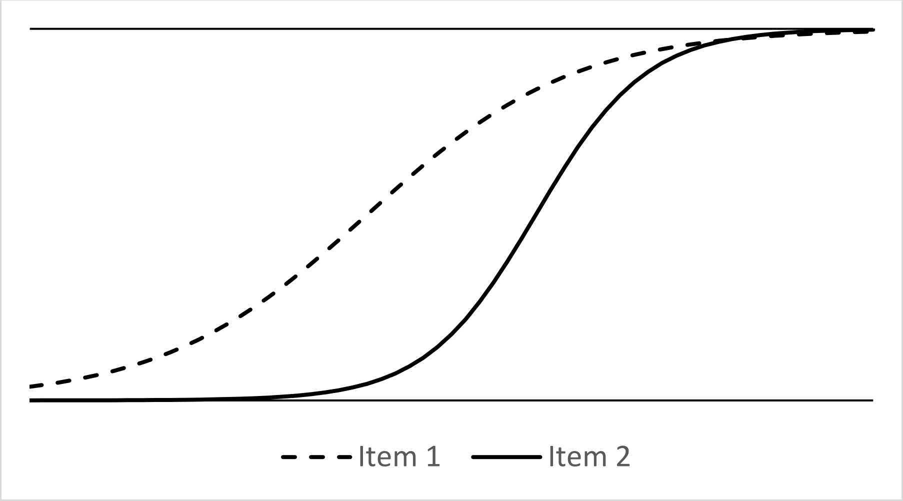

Figure 3. Item response curves in the two-parameter model.

The three-parameter model assumes that items can differ in difficulty, discrimination, and ease of guessing the correct answer (but not in any other way that affects any test taker's probability of answering correctly). Figure 4 shows the three-parameter response curves for two items. Item 2 has a higher difficulty parameter and a higher discrimination parameter than Item 1, but it is easier to guess. As you can see, the two items are about equally difficult for medium-ability and high-ability test takers, but Item 2 is easier for low-ability test takers.

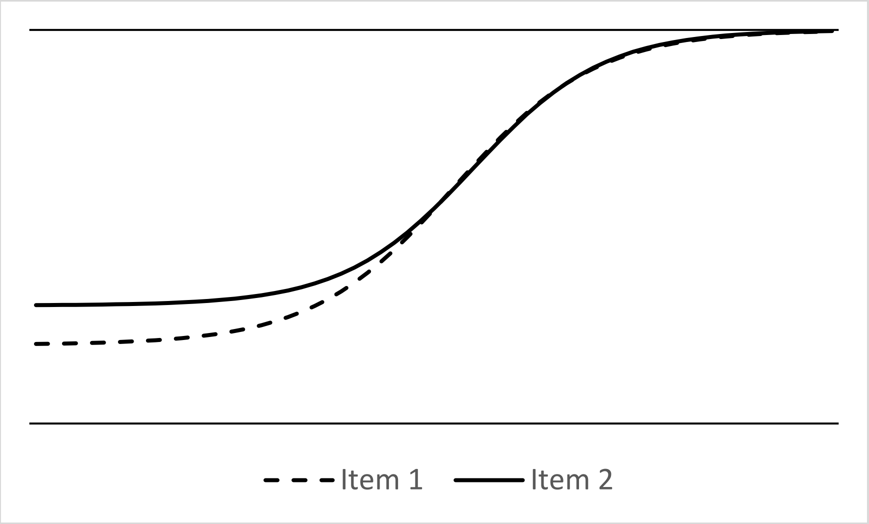

Figure 4. Item response curves in the three-parameter model.

Calibration

In IRT, calibration is the process of determining the response curves for a group of items measuring the same ability. For the one-parameter model, it means estimating a single number for each item—its difficulty. For the two-parameter model, it means estimating two numbers for each item—its difficulty and its discrimination. For the three-parameter model, it means estimating three numbers for each item—its difficulty, its discrimination, and the ease of guessing the correct answer.

To estimate these numbers, we need data from test takers taking the items. We don't need every test taker to take every item, but we do need data that connect all the items to be calibrated together. The connection between any two items can be direct or indirect. An item can be linked directly to another item if enough test takers have taken both those items. Two items that were not taken by the same test takers can be linked indirectly to each other, if they are linked to other items taken by the same test takers. Figure 5 is a diagram showing one possible pattern for linking items directly and indirectly. In this example, Items 1–10 are linked directly to Items 11–20, because test takers 1 to 97 took both those groups of items. Similarly, Items 1–10 are linked directly to Items 21–30 and to Items 31–40, because in each case, there is a group of test takers who took both groups of items. Items 11–20 are not linked directly to Items 21–30, because nobody took both those groups of items. However, both these groups of items are linked directly to Items 1–10. Therefore, Items 11–20 are linked indirectly to Items 21–30.

	Items 1–10	Items 11–20	Items 21–30	Items 31–40
Test takers 1–97	Data	Data		
Test takers 98–183	Data		Data	
Test takers 194–311	Data			Data

Figure 5. A data collection design linking four groups of items.

How large a data set is necessary for calibrating the items? There is no simple answer to this question. For statistical estimation, larger samples produce more accurate estimates. However, the more responses you already have, the less the improvement from adding another 100 responses. Calibration with the two-parameter model requires more data than calibration with the one-parameter model, and calibration with the three-parameter model requires even more. Most IRT experts would probably agree that an adequate calibration with the one-parameter model requires at least 150 responses to each item and that an adequate calibration with the two-parameter model requires at least 250. Many experts would argue for larger sample sizes than these.

A Calibrated Item Pool

A large group of items that have been calibrated together is called a “calibrated item pool” (or sometimes as an “item bank”). Having a calibrated item pool makes it possible to do some interesting and useful things. If we create a new form of the test from items in the pool, we can create a table that converts number-correct scores on that test form into scaled scores that are adjusted for difficulty. And we can do it before anybody ever takes that test form. The calibration gives us all the information we need. The scaled scores will be correct, if the response curves for the items in that test form have not changed since the items were calibrated—that is, if the test takers still respond to each item in the same way. (Of course, that is a big “if.”)

The table that converts number-correct scores on a new test form into scaled scores is called a “raw-to-scale conversion table.” If we can create it before anybody takes that new form, we can report scaled scores without waiting for the data from the new test form to be analyzed. We can calculate each test taker’s scaled score as soon as the test taker finishes the test.

Because we can create the raw-to-scale conversion table for a new form without waiting for test takers take that form, we can create a new test form for a small number of test takers. We can even create a test form by selecting items at random for each individual test taker, so that very few test takers will get the same selection of items. We can do all of these things by selecting items from a calibrated item pool.

Adding Items to the Pool

For most tests, we don’t want to limit all future forms of the test to the items originally in the pool. We want to add new items to the pool, so they can be used in future forms of the test. But we need to have those new items calibrated in a way that is consistent with the calibration of the items already in the pool. To do that, we need data to link the new items to the items already in the pool. The best way to get that kind of data is by a procedure called “embedded pretesting.”

“Pretesting” the new items means administering them to a sample of the test takers, but without using them to determine any test taker’s score. “Embedded” means that the new items being pretested are placed into a regular form of the test, along with the items that will determine the test takers’ scores. Of course, the test takers are not told which items are included in the scoring and which are being pretested. The test takers taking each new item provide a link between that item and the items used in computing their scores—items that are already in the calibrated item pool. This information is used to calibrate the new items—to estimate their response curves in a way that is consistent with the response curves of the items in the pool. Once the new items have been calibrated in this way, they can be added to the item pool.

Partial Credit Items

A "partial credit item" is an item with more than two possible scores. A test taker can earn full credit (the highest possible score), no credit (the lowest possible score), or partial credit (one or more scores in between). Constructed-response items (e.g., essay questions) are the most familiar type of partial credit item. Another type of partial credit item is the multiple-selection item, which is like a multiple-choice item with more than one correct choice. The test taker is instructed to select all correct choices, with partial credit for a response that is partially correct.

The IRT models for partial credit items estimate a response curve for each possible item score (except the lowest). If the possible item scores are 0, 1, 2, 3, and 4, there will be four response curves for that item. There will be a response curve for an item score of 1, another for an item score of 2, and so on. These IRT models allow for partial credit items to be calibrated together with 1/0 items (i.e., items on which the only possible scores are 1 and 0). When we calibrate both types of items together, we are implicitly assuming that they measure the same ability—that the test takers who tend to do well on the 1/0 items will also tend to do well on the partial credit items.

Determining Test Takers' Scores

There are two ways of determining test takers' scores on an IRT-based test: "number-correct scoring" and "pattern scoring." On tests that use the one-parameter IRT model, the two ways of scoring produce the same result, but on tests that use the two-parameter model (or the three-parameter model), they do not.

With number-correct scoring, the test taker's score on the test is computed from the number of items answered correctly. (If the test includes partial credit items, the number of 1/0 items answered correctly is combined with the number of points awarded for responses to the partial credit items.) For each form of the test, a conversion table translates each possible number-correct score into a scaled score. On tests that use IRT, the item response curves provide the information necessary to compute this table.

With pattern scoring, the relationship between the test taker's responses and the resulting score is more complex. It is possible for two test takers to answer the same number of items correctly and still get different scores. The test taker's score is based on an estimate of the test taker's ability, and that estimate depends on which specific items the test taker answered correctly.

The ability estimate for pattern scoring is usually computed according to a statistical principle called "maximum likelihood." The test taker's estimated ability is the ability value for which the test taker's pattern of responses has the highest probability of occurring. For example, suppose a test taker answers nearly all of the items correctly. That pattern of responses is highly probable if the test taker's ability is high. It is not very probable if the test taker's ability is only medium, and even less probable if the test taker's ability is low. Therefore, the maximum likelihood estimate of this test taker's ability will be a high value.

Pattern scoring can lead to problems when a test taker answers difficult items correctly but misses easy items. Such a pattern of responses would not be very probable for a high-ability test taker, but it also would not be very probable for a medium-ability test taker or a low-ability test taker. As a result, the maximum likelihood estimate may not be close to the test taker's actual ability.

Defining the Score Scale

There are two main approaches to defining the score scale for an IRT-based test. One way is to base the scaled score on the test taker's estimated ability. Tests scored by pattern scoring use this approach, and some tests scored by number-correct scoring use it also. The scaling procedure takes the test taker's estimated ability value, multiplies it by one number, and then adds another number. Those numbers are chosen to produce scaled scores that fit into a specified numerical range, such as 130 to 170.

The other main approach to defining the score scale is to specify a base form for the test. The base form can be any combination of items in the calibrated item pool. It does not have to be a test form that anybody ever takes. It can even be the entire item pool. The test taker's number-correct score is translated into an equivalent score on the base form. That base-form number-correct score is then translated to a scaled score, usually by multiplying it by one number and adding another.

These two different ways of defining the score scale have one consequence that sometimes confuses people. It has to do with the information the score gives us about test takers with extremely high scores. Think about a test taker who answers all the items correctly. If the score scale is based on the test taker's estimated ability, there is a lot of uncertainty in this test taker's score. We know that the test taker's ability is high, but we don't know how high. To find out, we would need a more difficult test. However, if the score scale is based on the test taker's number-correct score on a base form, there is very little uncertainty in the test taker's score. The test taker is likely to be strong enough to answer all the items on the base form (or on any other form made up of items from the pool).

Item Information

One application of IRT that is useful in selecting items for a new form of a test is called the "information function." It can be computed for an individual item, for a group of items, or for a complete test form, once the items have been calibrated onto the same ability scale. At any given point on the ability scale, the information function shows how strongly the item discriminates between the test takers above that point and the test takers below it.

Figure 6 shows the item response curves and the information functions for the three items in Figure 2. Item 1 is an easy item. It discriminates test takers in the low part of the ability range from those in the middle and high parts of the range. Its item response curve rises steeply in the lower middle part of the ability range, and that is where its item information function is highest. Item 2 is a medium-difficulty item. It discriminates test takers in the lower parts of the ability range from those in the upper parts of the range. Its item response curve rises steeply in the middle of the range, and that is where its item information function is highest. Item 3 is a difficult item, which discriminates test takers in the upper part of the ability range from those in the middle and lower parts of the range. Its item response curve rises steeply in the upper middle part of the ability range, and that is where its item information function is highest.

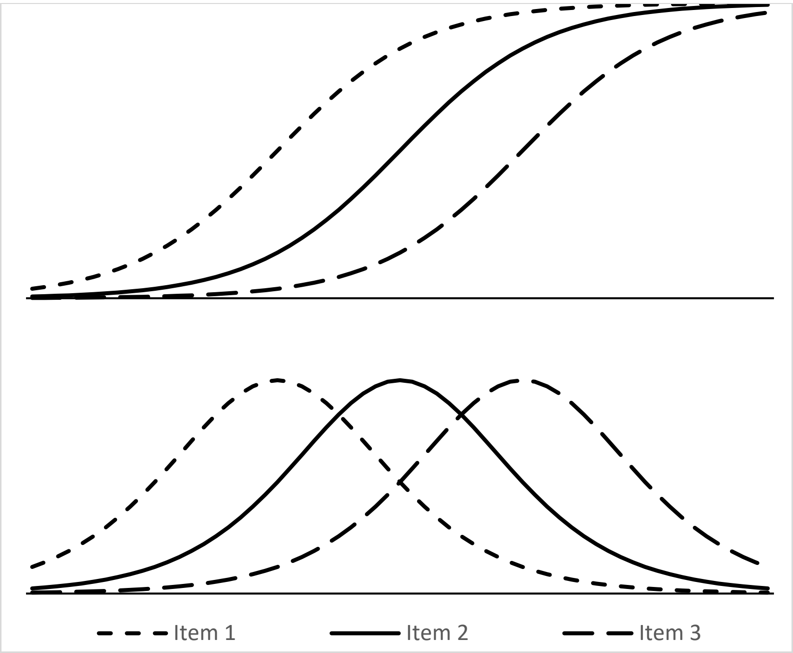

Figure 6. Information curves for the items in Figure 2.

Figure 7 shows the item response curves and the information functions for the two items in Figure 3. The steeper slope of the item response curve for Item 2 makes its information function higher than the information function for Item 1.

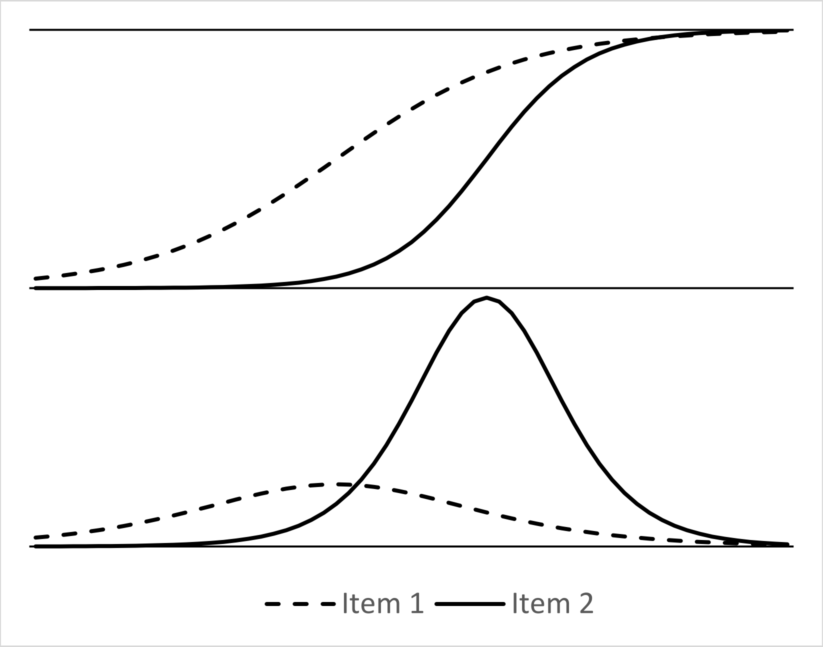

Figure 7. Information curves for the items in Figure 3.

The Test Information Function

The test information function is the sum of the information functions for all the items in the test. It shows how much information the test is providing about the test takers' ability at each point in the ability range. The more information, the better the estimate of the test takers' ability.

Each item that the test developer adds to the test will raise the test information function, but the increase will be greater at some points of the ability range and smaller at others. The higher the item information function is at any point in the ability range, the more the item will increase the test information function at that point.

If the test is used with a pass/fail cut point, the test developers will want the test information function to be high in the portion of the ability range near the cut point. They will meet that requirement by selecting items with item information functions that are high in that narrow range. If the test is intended to measure over a wide range of ability, the test developers will want the test information function to be fairly high over a wide range of ability. They will meet that requirement by selecting items with information functions that peak at many different ability levels.

At some testing organizations, the test developer assembling a new test form can use a computer program that compares the test information function with a "target information curve." Each time the test developer adds an item to the test form, the computer recomputes the test information function and displays it on a graph, along with the target information curve. By comparing the two curves, the test developer can see the parts of the ability range where more information is needed. The test developer can then select items that provide information at those ability levels.

The item information curves help the test developer choose items that provide discrimination in the parts of the ability range where it is needed. They do not help the test developer make sure the test meets the specifications for content—the specific types of knowledge and skills that the test is intended to measure. Two items that measure very different points of knowledge can have nearly identical response curves and information functions. An important part of the test developer's job is to make sure that the content of the items meets the specifications for the test. As far as IRT is concerned, "ability" is whatever the test measures. It is up to the test developer to make sure that the ability the test is measuring is what the test specifications call for.

References

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Appendix: Glossary

Ability: The characteristic of the test taker that the test measures (e.g., knowledge of U.S. history).

Calibration: The process of estimating the item response curves for a set of items.

Difficulty: The extent to which test takers tend *not* to respond correctly to an item.

Discrimination: The extent to which an item separates higher ability test takers from lower ability test takers.

Embedded pretesting: Collecting data about new items by inserting them into the operational test but not including them in the scores.

Item information function: A formula, table, or graph that indicates how much information the item provides about the ability of test takers at each point in the ability range.

Item pool: A collection of items available for creating a test.

Item response curve: A graph that shows the probability of a correct answer to the item, for test takers at each point in the ability range.

Maximum likelihood: In IRT, a statistical technique used for estimating a test taker's ability from the test taker's responses to the individual items on the test.

Model: In statistics, a set of mathematical statements that represents a real process, such as a test taker's response to a test item.

Parameter: In IRT, a number that indicates some characteristic of an item that affects the probability that test takers will respond correctly.

Partial credit item: An item with more than two possible item scores.

Pattern scoring: A type of scoring that takes into account which specific items the test taker answered correctly.

Probability: The number of times an event will occur in a particular set of circumstances, as a proportion of the number of times that set of circumstances occurs.

Test form: An edition of a test containing a particular set of items (i.e., not the same set of items as any other form of the test).

Test information function: A formula, table, or graph that indicates how much information the full test provides about the ability of test takers at each point in the ability range.

Unidimensionality: The assumption that all the items on the test measure the same characteristic of the test taker.

Notes

[1] IRT can be used to develop and score tests of psychological qualities that cannot reasonably be described as abilities, such as extraversion or compulsiveness. However, in the terminology of IRT, the quality that the test measures is called ability.