Suggested Readings

The classic presentation of generalizability theory is

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley.

The eminent psychometrician Lee Cronbach prepared a commentary on reliability theory and the widely used coefficient alpha nearly 50 years after its introduction. This commentary, as well as an argument that generalizability theory provides a more comprehensive perspective on reliability, is presented in

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386

A rather technical but clear and thorough discussion of generalizability theory is presented in

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

A widely used conceptual introduction to generalizability theory is presented in

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE.

A detailed description of using generalizability theory in personality assessment is given in

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *Journal of Personality Assessment*, *100*(1), 53–67. https://doi.org/10.1080/00223891.2017.1296455

A recent article articulating the value of generalizability theory and its connections to CTT and confirmatory factor analysis is

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018b). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. https://doi.org/10.1037/met0000107

The following chapter provides additional insight into variance components as reflecting consistency, which (as discussed throughout Chapters 5–7 of this book) is at the heart of reliability:

Furr, R. M. (2021). Variability and consistency in the analysis of behavior. In J. Rauthmann (Ed.), *Handbook of personality dynamics and processes* (pp. 733–761). Elsevier.

Notes

1. Here we use N – 1 in the denominator of the equation, rather than N.
2. Again these values are computed with N – 1in the denominator, rather than N.
3. These equivalences are true for variance components as estimated via ANOVA's mean square values (e.g., Tables 13.3 and 13.6). When estimated via different approaches, variance components will not exactly equal covariances as described in this section. However, even under different estimation approaches, they are typically very similar to the terms as outlined in this section.

Student Resources

This text includes access to datasets and files in R for the Technical Appendix. To learn more, visit **edge.sagepub.com/furr4e**.

14

ITEM RESPONSE THEORY AND RASCH MODELS

Item response theory (IRT) is another contemporary alternative to classical test theory (CTT). Although the roots of IRT have a long history (e.g., Lord, 1953; Rasch, 1960), IRT has emerged relatively recently as an alternative way of conceptualizing and analyzing measurement in the behavioral sciences. IRT is more computationally complex than CTT, but its proponents suggest that this complexity is offset by several important advantages.

This chapter presents an overview of IRT, with the goal of outlining the conceptual basis of IRT in a relatively accessible manner. In many areas of contemporary psychological testing, IRT is an increasingly central part of test development and evaluation. Thus, for readers wishing to be familiar with some of the most contemporary approaches to test development and psychometrics, IRT is an important topic. We won't discuss IRT in all its complexity and scope; indeed, IRT can be quite technically complex, and its scope is increasing consistently. Rather, the purpose of this chapter is to provide a solid overview of some of the most fundamental concepts of IRT and to do so in a way that will be relatively straightforward and conceptual.

We will discuss some of the most fundamental concepts in IRT, including the idea of item parameters, measurement models, and test information. We'll use examples to help clarify these concepts, hopefully enhancing readers' overall grasp of IRT. Finally, we will discuss several of the ways in which IRT has been used in psychological testing, demonstrating the breadth and applicability of this emerging psychometric approach.

FACTORS AFFECTING RESPONSES TO TEST ITEMS

At its heart, **item response theory (IRT)** is a psychometric approach emphasizing the fact that an individual's response to a particular test item is influenced by

qualities of the individual and by qualities of the item. Based on this perspective, IRT provides procedures for obtaining information about individuals, items, and tests. Advocates of IRT state that these procedures produce information that is superior to the information produced by CTT. As we will see, various forms of IRT exist, representing different degrees of complexity or different applicability to various kinds of tests. In this way, IRT is indeed flexible and broad.

Imagine that Suzy takes a five-item test of mathematical ability. According to the most basic form of IRT, the likelihood that Suzy will respond correctly to Item 1 on the test is affected by two things. The first is Suzy's mathematical ability. If Suzy has high mathematical ability, then she will have a relatively high likelihood of answering the item correctly. On the other hand, if Suzy has a low level of mathematical ability, then she will have a relatively low likelihood of answering the item correctly. The second thing affecting the likelihood that Suzy will respond correctly to Item 1 is that item's difficulty. If Item 1 is difficult, then Suzy will have a relatively low likelihood of answering the item correctly. If Item 1 is easy (i.e., has a low level of difficulty), then Suzy will have a relatively high likelihood of answering the item correctly. Therefore, the probability that Suzy will respond correctly to Item 1 is affected by her mathematical ability and by the difficulty of Item 1.

This logic can be extended to various kinds of psychological measures, but the basic form of IRT states that a person's response to an item is affected by the individual's trait level (e.g., Suzy's mathematical ability) and the item's difficulty level. As summarized in Figure 14.1 and the top half of Table 14.1, more complex forms of IRT include additional factors (or parameters) affecting a person's responses to items.

Respondent Trait Level as a Determinant of Item Responses

One factor affecting a person's probability of responding in a particular way to an item is the individual's level on the psychological trait being assessed by the item. For example, a person who has a high level of mathematical ability is more likely to

FIGURE 14.1 ⬢ Overview of Individual and (Some) Item Characteristics That Affect Responses to Test Items

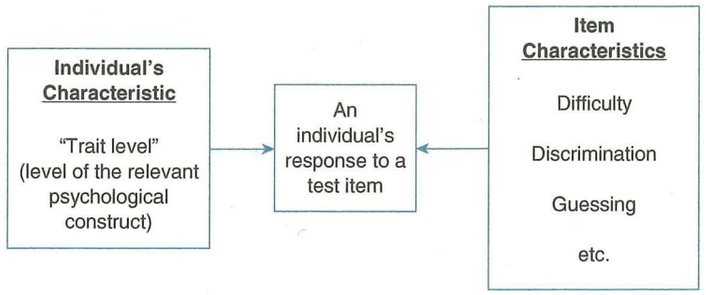


TABLE 14.1 ⬢ (Some) Factors Affecting Responses to Items and (Some) Widely Used IRT Models

(Some) Elements of Item Response Functions	Meaning
θ_s	An individual's trait level
β_i	An item's difficulty
α_i	An item's discrimination
c_i	Guessing value; probability of answering an item correctly, purely on the basis of chance
$P(X_{is}=1\|\ldots)$	The probability that a correct response will be made by a particular individual (s) when answering a particular item (i)
IRT Model	**Item Response Function**
1PL/Rasch	$P(X_{is}=1\|\theta_s,\beta_i)=\frac{e^{(\theta_s-\beta_i)}}{1+e^{(\theta_s-\beta_i)}}$
2PL	$P(X_{is}=1\|\theta_s,\beta_i,\alpha_i)=\frac{e^{(\alpha_i(\theta_s-\beta_i))}}{1+e^{(\alpha_i(\theta_s-\beta_i))}}$
3PL	$P(X_{is}=1\|\theta_s,\beta_i,\alpha_i,c_i)=c_i+(1-c_i)\frac{e^{(\alpha_i(\theta_s-\beta_i))}}{1+e^{(\alpha_i(\theta_s-\beta_i))}}$
GRM	$P(X_{is}\geq j\|\theta_s,\beta_{ij},\alpha_i)=\frac{e^{(\alpha_i(\theta_s-\beta_i))}}{1+e^{(\alpha_i(\theta_s-\beta_i))}}$

respond correctly to a math item than will a person who has a low level of mathematical ability. Similarly, a person who has a high level of extroversion is more likely to endorse or agree with an item that measures extroversion than will a person who has a low level of extroversion. An employee who has a high level of job satisfaction is more likely to endorse an item that measures job satisfaction than will an employee with a low level of job satisfaction. In the mathematical IRT models that we will discuss, an individual's (or subject's) trait level is indicated by a theta value (θ_s).

Item Difficulty as a Determinant of Item Responses

An item's level of **difficulty** is another factor affecting a person's probability of responding in a particular way. A math item that has a high level of difficulty will be less likely to be answered correctly than a math item that has a low level of difficulty (i.e., an easy item). For example, the item "What is the square root of 10,000?" is less likely to be answered correctly than is the item "What is 2 + 2?" Similarly, an extroversion item that has a high level of difficulty will be less likely to be endorsed than an extroversion item that has a low level of difficulty.

At first, the notion of "difficulty" might not be intuitive with regard to some psychological attributes. For example, it might be odd to think of a personality inventory as having "difficult" items, but consider these two hypothetical items that you might find on a questionnaire to measure extroversion: "I enjoy having conversations with friends" and "I enjoy speaking in front of large audiences." Assuming that these two items are validly interpreted as measures of extroversion, the first item is, in a sense, easier to endorse than the second item. That is, it is likely that more people would say that it is enjoyable to have a conversation with friends than to speak in front of a large audience. Similarly, in the context of job satisfaction, the statement "My job is okay" is likely "easier" to agree with than is the statement "My job is the best thing in my life."

Although they are separate issues in an IRT analysis, trait level and item difficulty are intrinsically connected. In fact, **item difficulty** is conceived in terms of trait level. Specifically, a difficult item requires a relatively high trait level to be answered correctly, but an easy item requires only a low trait level to be answered correctly. Returning to the two math test items, students might need to have a ninth-grade mathematical ability in order to have a good chance of answering a square root question correctly. In contrast, they might need only a second-grade mathematical ability to have a good chance of answering an addition question correctly. Thus, the item's difficulty can be described in terms of levels of the relevant trait (e.g., math ability).

The connection between trait level and difficulty might be particularly useful for understanding the concept of item difficulty in personality inventories or attitude surveys. Recall the extroversion items mentioned earlier: "I enjoy having conversations with friends" and "I enjoy speaking in front of large audiences." Arguably, the first item is easier than the second. Put another way, the first item requires only a low level of extroversion to be endorsed, but the second would seem to require a much higher level of extroversion to be endorsed. That is, even people who are fairly introverted (i.e., people who have relatively low levels of extroversion) would be likely to agree with the statement that they enjoy having conversations with their friends. In contrast, a person would probably need to be very extroverted to agree with the statement that they enjoy speaking in front of a large audience.

In an IRT analysis, trait levels and item difficulties are usually scored on a standardized metric (see Chapter 3), so that their means are 0 and the standard deviations are 1. Therefore, a person who has a trait level of 0 has an average level of that trait, and a person who has a trait level of 1.5 has a trait level that is 1.5 standard deviations above the mean. Similarly, an item with a difficulty level of 0 is an average item, and an item with a difficulty level of 1.5 is a relatively difficult item.

More specifically, for most examples that we will discuss, an item's difficulty is defined as the trait level required for participants to have a .50 probability of answering the item correctly. If an item has a difficulty of 0, then a person with an

average trait level (i.e., a person with a trait level of 0) will have a 50:50 chance of correctly answering the item. For that same item (i.e., an item with a difficulty of 0), a person with a high trait level (i.e., a trait level greater than 0) will have a higher chance of answering the item correctly, and a person with a lower trait level (i.e., a trait level less than 0) will have a lower chance of answering the item correctly. More generally, higher difficulty levels indicate that higher trait levels are required for participants to have a 50:50 chance of answering the item correctly. For example, if an item has a difficulty of 1.5, then a person with a trait level of 1.5 (i.e., a trait level that is 1.5 standard deviations above the mean) will have a 50:50 chance of answering the item correctly. Similarly, lower difficulty levels indicate that only relatively low trait levels are required for participants to have a 50:50 chance of answering the item correctly. In the mathematical IRT models that we will discuss, an item's difficulty level is indicated by a beta value (β_i, see Table 14.1).

Item Discrimination as a Determinant of Item Responses

Just as a test's items might differ in terms of their difficulties (with some items being more difficult than others), they might also differ in terms of the degree to which they can differentiate individuals who have high trait levels from individuals who have low trait levels. This item characteristic is called item discrimination, and it is analogous to an item–total correlation from CTT (Embretson & Reise, 2000).

An item's discrimination value indicates the relevance of the item to the trait being measured by the test. An item with a positive discrimination value is at least somewhat consistent with the underlying trait being measured, and a relatively large discrimination value (e.g., 3.5 vs. .5) indicates a relatively strong consistency between the item and the underlying trait. In contrast, an item with a discrimination value of 0 is unrelated to the underlying trait supposedly being measured, and an item with a negative discrimination value is inversely related to the underlying trait (i.e., high trait scores make it *less* likely that the item will be answered correctly). Thus, it is generally desirable for items to have a large positive discrimination value.

Why would some items have good discrimination and others have poor discrimination? Consider the following two items that might be written for a math test:

1. How many pecks are in three bushels? (a) 12 (b) 24
2. What is 10 times 10? (a) 10 (b) 100

Consider the first item. What is required of a respondent to answer this item correctly? To answer the item correctly, the respondent needs to have enough mathematical ability to perform multiplication. However, this item also requires knowledge of the number of pecks in a bushel. The fact that this item requires something apart from basic mathematical ability means that it is not very closely—or at least

not purely—related to mathematical ability. In other words, having a high level of mathematical ability is not enough to answer the item correctly. The student might have the ability to multiply 4 times 3; however, they might not have a very good chance of answering the item correctly without also knowing that there are four pecks in a bushel. Thus, this item would likely have a low discrimination value, as it is only weakly related to the underlying trait being assessed by the test of mathematical ability. In other words, this item does not do a very good job of discriminating students who have a relatively high level of mathematical ability from those who have relatively low mathematical ability. Even if Suzy answers the item correctly and Johnny answers the item incorrectly, we might not feel confident concluding that Suzy truly has a higher level of mathematical ability than does Johnny—perhaps Johnny has the mathematical ability, but he simply does not know the number of pecks in a bushel (see our discussion of "construct-irrelevant content" in Chapter 8).

Now consider the second math item. What is required of a respondent to answer it correctly? This item requires the ability to perform multiplication, but it requires no additional knowledge or ability. The only quality of the student that is relevant to answering the item correctly is mathematical ability. Therefore, it is a much more "pure" mathematical item, and it is likely more strongly related to the underlying trait of mathematical ability than is the first item. Consequently, it would likely have a relatively high discrimination value. In other words, this item does a better job of discriminating individuals who have a relatively high level of mathematical ability from those who have relatively low mathematical ability. If Suzy answers the item correctly and Johnny answers the items incorrectly, then we feel fairly confident concluding that Suzy has a higher level of mathematical ability than does Johnny. In the mathematical IRT models that we will discuss, an item's discrimination level is indicated by an alpha value (α_i, see Table 14.1).

Guessing

Guessing is related to a third item property that might affect participants' responses to some types of test items. For tests such as multiple-choice exams or true/false exams of knowledge or ability, some test takers might resort to guessing if they do not know the correct answer to an item. When taking such tests, participants might answer some items correctly, simply on the basis of guessing and chance.

Thus, IRT can include a guessing component to account for this possibility, and it is related to the probability that participants will answer an item correctly purely on the basis of chance. For example, the guessing component for true/false items accounts for the fact that guessing will produce a correct answer 50% of the time. Similarly, the guessing component for multiple-choice items depends on the number of response options that are available—if there are four response options, then the guessing component accounts for the fact that guessing will produce a correct answer 25% of the time. This property is mainly relevant to tests of knowledge, skill,

ability, or achievement, rather than personality or attitudes. That is, it is relevant for items that are scored as correct or incorrect; thus, it is likely to be examined only for tests with those types of items. In the mathematical IRT models that we will discuss, an item's guessing component is indicated by c_i (see Table 14.1).

IRT MEASUREMENT MODELS

From an IRT perspective, we can identify the components (e.g., difficulty, discrimination, etc.) affecting the probability that a person will respond in a particular way to a particular item. A *measurement model*, or *item response function*, expresses the links between an outcome (e.g., a person's response to a particular item) and the components that affect that outcome (e.g., qualities of the person and/or qualities of the item). IRT models express these links mathematically, in terms of the probability that a person with a particular trait level will respond in a particular way to a particular item.

As summarized in Table 14.2, a variety of models have been developed from the IRT perspective, and these models differ from each other in at least two important ways (note that Table 14.2 reflects only a small portion of the many IRT models, and it does not include all of the ways in which models might differ from each other). One important difference among the measurement models is in terms of the item characteristics, or *parameters*, that are included in the models. That is, some models are designed to account for only one parameter, whereas other, more complex models account for two or more parameters. A second important difference among measurement models is in terms of the response option format. For example, some models are designed to be used for dichotomous (or binary) items, such as true/false, yes/no, or correct/incorrect items. Others are designed for items with more than two response options (i.e., polytomous items), such as Likert-type items with five options reflecting one's level of agreement.

TABLE 14.2 ⬢ (Some) IRT Models, in Terms of Their Parameters and Their Response Option Format

Parameters That Vary From Item to Item	Response Option Format	
	Dichotomous	Polytomous
Difficulty	Rasch model, 1PL	Partial credit model, rating scale model
Difficulty and discrimination	2PL	Graded response model, generalized partial credit model, nominal response model
Difficulty, discrimination, and guessing	3PL	

This section describes four of the more commonly used IRT measurement models. These descriptions are intended to convey the basic logic and meaning of the models, to further illuminate the logic and nature of an IRT perspective on psychometrics. These models are summarized in the bottom of Table 14.1, presented earlier.

One-Parameter Logistic Model (or Rasch Model)

The simplest IRT model is often called the **Rasch model** or the **one-parameter logistic model (1PL)**. According to this model, a person's response to a dichotomous item (i.e., right/wrong, true/false, agree/disagree) is determined by the individual's trait level and only a single item characteristic or parameter—the item's difficulty.

One way of expressing the Rasch model is in terms of the probability that a person with a particular trait level will correctly answer an item that has a particular difficulty. The item response function for the Rasch model is often presented as follows (e.g., Embretson & Reise, 2000):

$$P(X_{is}=1|\theta_s,\beta_i)=\frac{e^{(\theta_s-\beta_i)}}{1+e^{(\theta_s-\beta_i)}}. \qquad (14.1)$$

This equation might require some explanation:

P refers to probability. In this case, it refers to a "conditional" probability, as we will describe.

X_{is} refers to a particular response (X) made by individual (or subject) s to item i. More specifically, "$X_{is}=1$" refers to a "correct" response or an endorsement of the item by subject s.

θ_s refers to the trait level of subject s.

β_i refers to the difficulty of item i.

e is the base of the natural logarithm (i.e., $e=2.7182818\ldots$), found on many calculators.

So $P(X_{is}=1|\,\theta_s,\beta_i)$ refers to the probability (P) that subject s will respond to item i correctly. The vertical bar in this statement indicates that this is a "conditional" probability. That is, it indicates that the probability that the subject will correctly respond to the item depends on (i.e., is conditional upon) the subject's trait level (θ_s) and the item's difficulty (β_i). As mentioned earlier, in an IRT analysis, trait levels and item difficulties are usually scaled on a standardized metric, so that their means are 0 and the standard deviations are 1.

Consider two examples, in terms of a math test. First, what is the probability that a person who has an above-average level of math ability (say, a level of math ability that is 1 standard deviation above the mean, $\theta_s=1$) will correctly answer an item that has a relatively low level of difficulty (say, $\beta_i=-.5$)? Entering these parameter values into the 1PL equation (Equation 14.1), we have

$$P(X_{is}=1|1,-.5)=\frac{e^{(1-(-.5))}}{1+e^{(1-(-.5))}}=\frac{e^{(1.5)}}{1+e^{(1.5)}}=\frac{4.48}{1+4.48}=.82.$$

These calculations indicate that there is a .82 probability that the individual will correctly answer the item. In other words, there is a high likelihood (i.e., greater than an 80% chance) that this individual will answer correctly. This should make intuitive sense, because a person with a *high* level of ability is responding to a relatively *easy* item.

Now, what is the probability that a person who has a below-average level of math ability (say, a level of math ability that is 1.39 standard deviations below the mean, $\theta_s=-1.39$) will correctly answer an item that has a relatively low level of difficulty (say, $\beta_i=-1.61$)? Again entering these values in the 1PL model (Equation 14.1), we have

$$P(X_{is}=1|-1.39,-1.61)=\frac{e^{(-1.39)-(-1.61)}}{1+e^{(-1.39)-(-1.61)}}=\frac{e^{(.22)}}{1+e^{(.22)}}=\frac{1.25}{1+1.25}=.56.$$

This indicates that there is a .56 probability that the individual will correctly answer the item. In other words, there is slightly more than a 50:50 chance that this individual will answer correctly. This should make intuitive sense because the individual's trait level ($\theta=-1.39$) is only slightly higher than the item's difficulty level ($\beta=-1.61$). Recall that the item difficulty level represents the trait level at which a person will have a 50:50 chance of correctly answering the item. In the current example, the individual's trait level is slightly higher than the item's difficulty level, so the probability that the individual will correctly answer the item is slightly higher than .50.

In sum, Equation 14.1 represents the 1PL model. This model is intended to reflect the connections between respondent trait level, item difficulty level, and the probability of correctly answering a dichotomous (binary) item.

Such models or functions are the foundations of IRT, and they lay the ground for important aspects of measurement and psychometrics. For example, based on the mathematical connections defined in such models, and based on a set of actual responses (from multiple test takers) to a set of items, we can estimate values for the parameters in the model. That is, we can use specialized software to estimate a trait level for each person and a difficulty level for each item (note that we simply assumed knowledge of those values for the purpose of the illustrative examples above). Indeed, this is how some contemporary high-stakes tests are scored. In addition, once we know (or at least once we have estimated) respondents' trait levels and the items' difficulty levels, we can estimate the probability that each respondent will correctly answer each item. Such information is at the root of an IRT-based psychometric evaluation of tests and their items. We will cover these

issues in detail later—scoring, estimation of item parameters, psychometric evaluation of tests, and so on.

The purpose of outlining the 1PL and other such models (discussed next) at this point is to convey some foundational ideas of IRT. That is, to understand how IRT-based procedures can be used to create scores, evaluate items, design tests, and so on, it is important to understand the basic idea that there are proposed mathematical connections between specific characteristics of items, characteristics of people, and the likelihood of responses. Once this basic idea is clear, and once you have a sense of the various models that have been proposed, the use and implications of IRT will become clearer.

Returning to the 1PL, it is a relatively simple model. It applies only to dichotomous items and includes only one item parameter—item difficulty. As shown earlier in Table 14.2, other IRT measurement models are more complex, in that they include additional item parameters. Such models might be better than the 1PL, in terms of estimating respondents' trait levels and respondents' likelihood of answering each item correctly.

Two-Parameter Logistic Model

A slightly more complex IRT model is called the **two-parameter logistic model (2PL)** because it includes two item parameters. According to the 2PL model, a person's response to a binary item is determined by the individual's trait level, the item difficulty, and the item discrimination. Thus, the difference between the 2PL and the 1PL, or Rasch model, is the inclusion of the item discrimination parameter. This can be presented as follows (e.g., Embretson & Reise, 2000):

$$P(X_{is}=1\,|\,\theta_s,\beta_i,\alpha_i)=\frac{e^{(\alpha_i(\theta_s-\beta_i))}}{1+e^{(\alpha_i(\theta_s-\beta_i))}}, \qquad (14.2)$$

where α_i refers to the discrimination of item i, with higher values representing more discriminating items (and all other terms are defined as in the 1PL model described earlier). Thus, the 2PL model states that the probability of a respondent answering an item correctly is conditional on (i.e., determined by) the respondent's trait level (θ_s), the item's difficulty (β_i), and the item's discrimination (α_i).

Consider again the items "How many pecks are in three bushels?" and "What is 10 times 10?" Let's assume that the two items have equal difficulty in terms of the mathematical ability required to have a 50:50 chance of answering correctly (say, $\beta = -.5$). However, let's also assume that they have different discrimination values, as discussed earlier (say, $\alpha_1 = .5$ and $\alpha_2 = 2$).

Consider two potential responses to the first item. What is the probability that Suzy, who has an above-average level of math ability (say, a level of math ability that is 1 standard deviation above the mean, $\theta = 1$), will correctly answer Item 1?

Entering Suzy's trait level and the two item parameters in the 2PL model (Equation 14.2), we see that Suzy has a 68% chance of answering Item 1 correctly:

$$P(X_{is}=1\,|\,1,-.5,.5)=\frac{e^{(.5(1-(-.5)))}}{1+e^{(.5(1-(-.5)))}}=\frac{e^{(.75)}}{1+e^{(.75)}}=\frac{2.12}{1+2.12}=.68.$$

Now, what is the probability that Johnny, who has an average level of math ability ($\theta = 0$), will correctly answer Item 1? Again using the 2PL, we see that Johnny's likelihood is somewhat less, at 56%:

$$P(X_{is}=1\,|\,0,-.5,.5)=\frac{e^{(.5(0-(-.5)))}}{1+e^{(.5(0-(-.5)))}}=\frac{e^{(.25)}}{1+e^{(.25)}}=\frac{1.28}{1+1.28}=.56.$$

Take a moment to note the difference. Suzy's level of mathematical ability is substantially higher than Johnny's (i.e., $\theta = 1$ vs. $\theta = 0$), but her probability of answering the item correctly is only slightly higher than his (i.e., .68 vs. .56). This is a relatively large difference in trait level (1 entire standard deviation) but a relatively small difference in the likelihood of answering the item correctly. Thus, this item doesn't seem to reflect the difference between the two respondents very well—the people are quite highly different in terms of trait level, but they are not very different in terms of the likelihood that they will answer the item correctly. The difference in their trait levels is not clearly reflected by the difference in their probabilities for this item. That is, the item does not discriminate very robustly between people at different trait levels.

Now, let us consider the second item. Recall that Item 2 had a larger discrimination value, at $\alpha_2 = 2$. What are the probabilities that Suzy and Johnny will answer Item 2 correctly? Again using the 2PL, we obtain the following results:

$$\text{Suzy}: P(X_{is}=1\,|\,1,-.5,2)=\frac{e^{(2(1-(-.5)))}}{1+e^{(2(1-(-.5)))}}=\frac{e^{(3)}}{1+e^{(3)}}=\frac{20.09}{1+20.09}=.95.$$

$$\text{Johnny}: P(X_{is}=1\,|\,0,-.5,2)=\frac{e^{(2(0-(-.5)))}}{1+e^{(2(0-(-.5)))}}=\frac{e^{(1)}}{1+e^{(1)}}=\frac{2.72}{1+2.72}=.73.$$

Take a moment to consider the difference for Item 2. For this item, Suzy has a .95 probability of answering the item correctly, and Johnny has only a .73 probability of answering the item correctly. Recall that the difference between the students' trait level (i.e., mathematical ability) is 1 standard deviation; however (unlike for Item 1), Suzy's probability of answering Item 2 correctly is noticeably higher than Johnny's. Previously, for Item 1, there was only a .12 difference in their probabilities of answering correctly (.68 vs. .56). However, for Item 2, there is a more dramatic .22 difference in their probabilities of answering correctly (.95 vs. .73). Thus, as compared with Item 1, we see that Item 2—the item with the higher discrimination value—draws a more robust distinction between individuals who have different

trait levels. This type of information might be used to select items for a test, opting for items with greater discrimination values.

Just as the 2PL model is an extension of the Rasch model (i.e., the 1PL model), other models are extensions of the 2PL model. You might not be surprised to learn that the **three-parameter logistic model (3PL)** adds yet another item parameter. We turn to that next.

Three-Parameter Logistic Model

Earlier, we mentioned that guessing is an issue that might affect responses to some types of tests. For example, multiple-choice tests of knowledge provide test takers with a limited number of potentially correct answers, and respondents must choose which is the correct one.

For such tests, even if an individual randomly guesses among the choices, she or he is likely to choose a correct response some proportion of the time. Consider a multiple-choice item with four response options. Someone who simply randomly chooses one of the four options has a .25 probability of being correct. Thus, .25 would be the lowest probability that would be expected for answering that item correctly. That is, even someone with an extraordinarily low trait (or knowledge) level would have at least a .25 probability of choosing the correct answer for a multiple-choice item with four options. Of course, as a respondent's trait level increases, his or her probability of choosing the correct answer begins to increase above a pure "random chance" level of .25. Thus, .25 is the "lower bound" or "lower asymptote" of the response probabilities for the item.

In terms of an IRT perspective, then, a lower-bound adjustment for guessing can be incorporated into a measurement model. The three-parameter logistic model (or 3PL) does just that. The 3PL can be seen as a variation on the 2PL, discussed earlier (Equation 14.2). The 3PL is

$$P(X_{is}=1\,|\,\theta_s,\beta_i,\alpha_i,c_i)=c_i+(1-c_i)\frac{e^{(\alpha_i(\theta_s-\beta_i))}}{1+e^{(\alpha_i(\theta_s-\beta_i))}}, \tag{14.3}$$

where c_i refers to the lower-bound probability of correctly answering item i (and all other terms are defined earlier as in the 1PL and 2PL model). Thus, the 3PL model states that the probability of a respondent answering an item correctly is conditional on (determined by) the respondent's trait level (θ_s), the item's difficulty (β_i), the item's discrimination (α_i), and a "guessing parameter" representing the likelihood of correctly answering the item purely on the basis of guessing. As a technical aside, the difficulty parameter (β_i) in the 3PL is no longer the trait level at which someone has a .50 probability of correctly answering the item; rather, it is now the trait level at which someone has a (1 + c)/2 chance of correctly answering

the item—the midpoint between the probability based on randomly guessing and a 1.0 probability.

To illustrate the 3PL, let's imagine an individual taking a multiple-choice history exam, with the following question: "Who was the first person to walk on the moon?" And imagine that the options were (a) Neil Armstrong, (b) Buzz Aldrin, (c) Yuri Gagarin, and (d) John Glenn. Let's imagine that this item has a difficulty of, say, $\beta = -.5$ and a discrimination of $\alpha = 1.2$. Moreover, because there are four response options, let's assume that the guessing parameter is $c = .25$ (i.e., $1/4 = .25$).

What is the probability that an individual with a relatively low level of "knowledge of history," say $\theta = -2.5$, would choose the correct answer?

$$P(X_{is}=1\,|\,-2.5,.5,1.2,.25)=.25+(1-.25)\frac{e^{(1.2(-2.5-.5)}}{1+e^{(1.2(-2.5-.5)}},$$

$$P(X_{is}=1\,|\,-2.5,.5,1.2,.25)=.25+(.75)(.027),$$

$$P(X_{is}=1\,|\,-2.5,.5,1.2,.25)=.25+.02,$$

$$P(X_{is}=1\,|\,-2.5,.5,1.2,.25)=.27$$

Based on the 3PL and these parameter values, the individual would have about a 27% chance of choosing the correct answer. Note that this is slightly better than purely guessing. Without the adjustment for guessing (i.e., using the 2PL), the probability would be only .027 (i.e., 2.7%). Hopefully it is clear that such a probability (close to a zero chance of choosing the correct answer) does not make sense in the context of a multiple-choice item with only four possible options. And hopefully this demonstrates the logic of the 3PL.

In sum, the 1PL, 2PL, and 3PL models are IRT measurement models that differ with respect to the number of item parameters that are included in the models. However, all of these models are designed for items with a dichotomous response outcome (e.g., correct vs. incorrect or agree vs. disagree). Even for multiple-choice tests of knowledge, which initially offer multiple possible answers, the ultimate response outcome is dichotomous—that is, is the chosen answer correct or incorrect? As mentioned earlier, there is at least one additional way in which IRT measurement models differ from each other—in terms of the response option format, particularly in terms of the number of response options.

Graded Response Model

Many tests, questionnaires, and inventories in the behavioral sciences include more than two response options. That is, many tests do not boil down to a dichotomous distinction between right and wrong, true and false, or agree and disagree.

For example, many personality questionnaires include self-relevant statements (e.g., "I enjoy having conversations with friends"), and respondents are given three or

more response options (e.g., *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*). Such items are known as **polytomous items**, and they require IRT models that are different from those required by binary items. Models such as the **graded response model (GRM)** (Samejima, 1969), the partial credit model (Masters, 1982), and the nominal response model (Thissen et al., 2010) are polytomous IRT models.

Although the models for polytomous items differ in terms of the response options that they can accommodate, they rely on the same general principles as the models designed for dichotomous items. That is, they reflect the idea that a person's response to an item is determined by the individual's trait level and by item properties, such as difficulty and discrimination.

However, these models are also more complex than the models for dichotomous items. For example, the GRM produces several difficulty parameters for each item, whereas (as noted earlier) the previous models include only a single difficulty parameter for an item. Furthermore, the nominal response model includes several discrimination parameters for each item.

We will briefly discuss the GRM for two reasons. First, it provides a useful example representing these more complex models, thus expanding the possible insights into IRT. A well-rounded basic understanding of IRT should include some familiarity with the logic of models designed for polytomous items, and the GRM is a good vehicle for doing so. Second, the GRM might be the most commonly applied IRT model in some areas of psychology, particularly in the measurement of personality and psychopathology (Preston et al., 2011). Thus, some readers might be particularly likely to encounter this model in many applications of IRT.

In a sense, the GRM initially conceptualizes a polytomous item in terms of several dichotomous response distinctions. Generally speaking, if there are m response options or categories, then there are $m - 1$ distinctions. For example, consider again a personality questionnaire item such as "I enjoy having conversations with friends," which might have five response options (e.g., *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*). For such an item, there are four possible distinctions: (1) the difference between *strongly disagreeing* with the item and *disagreeing*, (2) the difference between *disagreeing* and being *neutral*, (3) the difference between being *neutral* and *agreeing*, and (4) the difference between *agreeing* and *strongly agreeing*.

According to the GRM, each of these dichotomous differences can be represented as follows:

$$P(X_{is} \geq j \mid \theta_s, \beta_{ij}, \alpha_i) = \frac{e^{(\alpha_i(\theta_s - \beta_i))}}{1 + e^{(\alpha_i(\theta_s - \beta_i))}}, \tag{14.4}$$

where j refers to a particular response option, βij is a difficulty parameter for response option j on item i, and the other parameters are as defined earlier. $P(X_{is} \geq j \mid \theta_s, \beta_{ij}, \alpha_i)$ refers to the probability that a person with trait level s will

respond to item i in response option j or higher (which is conditional on the person's trait level, the difficulty of the relevant response distinction for category j of item i, and the discrimination value of item i). Because there are $m - 1$ dichotomous response differences, there are $m - 1$ difficulty parameters (β_{ij}s) for each item. Note, however, that the GRM includes only a single discrimination parameter (α_i) for an item.

Imagine that we use our statistical software to analyze responses to the item "I enjoy having conversations with friends" and find the following (estimated) difficulty parameters: $\beta_1 = -1.78$, $\beta_2 = -.57$, $\beta_3 = .45$, $\beta_4 = 1.53$, along with a discrimination parameter of $\alpha = 2.32$. The four β parameters indicate the difficulty associated with each response option dichotomy. That is, each difficulty value represents the trait level (i.e., θ) required to move from one response option to the next "higher" one on the scale. For example, the first difficulty parameter (β_1) reveals that someone who has a trait level of −1.78 has a 50% chance of responding higher than *strongly disagree* (i.e., responding *disagree* or more positively). Similarly, β_4 indicates that someone with a trait level of 1.53 will have a 50% chance of responding higher than *agree* (i.e., responding *strongly agree*).

Of course, response probabilities can be computed for people with particular trait levels, representing the probability that the person will respond in each of the response categories. For example, a person with an average level of extroversion ($\theta_s = 0$), which is presumably the trait that drives responses to the "I enjoy having conversations with friends" item, would have the following response probabilities:

- Responding *disagree* or higher (i.e., responding higher than *strongly disagree*):

$$P(X_{is} \geq \textit{disagree} \mid 0, -1.78, 2.32) = \frac{e^{(2.32(0-(-1.78)))}}{1 + e^{(2.32(0-(-1.78)))}} = \frac{62.15}{1 + 62.15} = .98$$

- Responding *neutral* or higher (i.e., responding higher than *disagree*):

$$P(X_{is} \geq \textit{neutral} \mid 0, -.57, 2.32) = \frac{e^{(2.32(0-(-.57)))}}{1 + e^{(2.32(0-(-.57)))}} = \frac{3.75}{1 + 3.75} = .79.$$

- Responding *agree* or higher (i.e., responding higher than *neutral*):

$$P(X_{is} \geq \textit{agree} \mid 0, .45, 2.32) = \frac{e^{(2.32(0-(.45)))}}{1 + e^{(2.32(0-(.45)))}} = \frac{.35}{1 + .35} = .26.$$

- Responding *strongly agree* (i.e., responding higher than *agree*):

$$P(X_{is} = \textit{strongly agree} \mid 0, 1.53, 2.32) = \frac{e^{(2.32(0-(1.53)))}}{1 + e^{(2.32(0-(1.53)))}} = \frac{.03}{1 + .03} = .03.$$

Take a moment to note the pattern of these probabilities, specifically the fact that they become smaller as the response options become more extremely positive (i.e., as the options reflect more "difficult" choices). We see that a person with an average trait level is very likely to respond higher than *strongly disagree*, which should make

intuitive sense. That is, a person with an average level of extroversion would be very likely to say something besides *strongly disagree* when asked if they enjoy having conversations with friends. Going further, we see that this person would have a nearly 80% chance of responding without any level of disagreement (i.e., responding *neutral* or higher). Going even further, we see that the probabilities become smaller and smaller as the options reflect greater degrees of endorsement of the item (i.e., as they reflect more extreme levels of extroversion).

Using these values, we can then estimate the probability that a person will choose a *specific* response option when responding to this item. That is, Equation 14.4 can be seen as reflecting a person's probability of responding in particular *ranges* of response options (e.g., responding *disagree or higher*, responding *agree or higher*, etc.). These values can now be used to find the likelihood that the person will make a *specific* response to the item (e.g., *disagree*, *agree*, *strongly agree*, etc.). This can be done by computing the difference between two adjacent "range" probabilities, as just discussed:

$$P(X_{is} = j \mid \theta_s, \beta_{ij}, \alpha_i) = P(X_{is} \geq j-1 \mid \theta_s, \beta_{ij}, \alpha_i) - P(X_{is} \geq j \mid \theta_s, \beta_{ij}, \alpha_i). \quad (14.5)$$

In this equation, j refers to one response option (e.g., *disagree*), and $j-1$ refers to the immediately prior option (e.g., *strongly disagree*). Thus, based on the probabilities computed earlier from Equation 14.4, we can use Equation 14.5 to estimate the following probabilities for a person with an average level of extroversion ($\theta_s = 0$):

- Probability of endorsing the *strongly disagree* option = 1 − .98 = .02 (note that the probability of responding in the range of strongly disagree or higher is 1.0, since that range includes the entire set of response options).
- Probability of endorsing the disagree option = .98 − .79 = .19.
- Probability of endorsing the neutral option = .79 − .26 = .53.
- Probability of endorsing the agree option = .26 − .03 = .23.
- Probability of endorsing the strongly agree option = .03 − .00 = .03 (note that the probability of responding higher than strongly agree is .00, since there are no response options beyond *strongly agree*).

Based on these values, the person is most likely to respond "neutral" to this item. The neutral response option has the largest probably (.53), as compared to all the other response options. Thus, it is the most likely response, according to the GRM.

Of course, people who have different trait levels will have different probabilities of responding to the various options. People with low trait levels will be relatively likely to respond with the lower response options (e.g., *strongly disagree* or *disagree*), indicating low endorsement of the trait. In contrast, people with high trait levels will be relatively likely to respond with higher response options, such as *agree* or *strongly agree*. By using the equations above, we could use the GRM to estimate response probabilities for any particular trait level.

OBTAINING PARAMETER ESTIMATES: A 1PL EXAMPLE

You might wonder how we actually obtain the estimates of trait level, item difficulty, and item discrimination that are entered into the equations described in the previous section. Purely for the sake of those earlier illustrations, we acted as if we already had information about trait levels, difficulty, and discrimination.

In real-world research and application, however, we would not have such independent knowledge. In fact, obtaining such information is precisely one of the main reasons that we might conduct an IRT-based analysis in the first place. That is, we might conduct an IRT-based analysis in order to obtain estimates of individuals' trait levels and items' difficulty or discrimination values.

This type of information is almost always estimated by using specialized statistical software to analyze individuals' responses to sets of items. Indeed, software packages such as PARSCALE, BILOG, SAS's "irt" procedure, Mplus, and R packages such as "ltm" and "mirt" allow researchers to conduct IRT-based analyses. As with other multivariate statistical procedures, the actual calculation of IRT-based parameter estimates is computationally complex. Nevertheless, an example of a relatively simple IRT analysis conducted "by hand" might give you a deeper sense of how the process works and thus give you a deeper understanding of IRT in general.

Table 14.3 presents the (hypothetical) responses of six individuals to five items on a test of mathematical ability. In these data, "1" represents a correct answer and "0" represents an incorrect answer. Such a small data set is not representative of "real-world" use of IRT. Ideally, we would have a very large data set, with many respondents and many items. However, a small data helps to illustrate IRT analysis as simply as possible.

An important step in IRT analysis is to choose an appropriate measurement model. Note that the responses in our example represent a binary outcome—correct versus incorrect. Therefore, we would choose a model that is appropriate for binary outcomes (e.g., the 1PL, 2PL, 3PL). Having focused on this class of models, we would then choose a model that includes parameters in which we are interested. An advanced issue involves an evaluation of which model "fits" best (see the next section). That is, we could conduct analyses to determine whether a particular model *should be* applied to a particular data set. At this point, however, we will use the Rasch model (the 1PL model) as the measurement model for analyzing these data because it is the simplest model.

Several kinds of information can be obtained from these data. Recall from our earlier description of the 1PL (Equation 14.1) that this model includes two determinants of an item response—the respondent's trait level and the items' difficulty level. Let's focus first on information about the respondents, and we will estimate a trait level for each of the six individuals who have taken the test. We will then estimate the item difficulties. As explained later, estimation usually proceeds through a series of

TABLE 14.3 ⬢ Raw Data for IRT Example: A Hypothetical Five-Item Test of Mathematical Ability

Person	Item 1	Item 2	Item 3	Item 4	Item 5
1	1	0	0	0	0
2	1	1	0	1	0
3	1	1	1	0	0
4	1	1	0	1	0
5	1	1	1	0	1
6	0	0	1	0	0

Note: IRT = item response theory.

iterative steps in which trait parameters and item parameters are initially estimated, then revised repeatedly until the estimates seem to be as accurate as possible.

The *initial* estimates of trait levels are obtained through a two-step process. First, we determine the proportion of items that each respondent answered correctly. For a respondent, the proportion correct is simply the number of items that they answered correctly divided by the total number of items that they answered. As shown in Table 14.3, Respondent 5 answered four of the five items correctly (4/5), so her proportion correct is .80. Table 14.4 presents the proportion correct for each respondent. To obtain initial estimates of trait levels, we take the natural log of the ratio of proportion correct to proportion incorrect:

$$\theta_s = \text{In}\left(\frac{P_s}{1-P_s}\right), \tag{14.6}$$

where P_s is the proportion correct for Respondent 5. This analysis suggests that Respondent 5 has a relatively high trait level:

$$\theta_s = \text{In}\left(\frac{.80}{1-.80}\right) = \text{In}(4) = 1.39.$$

This initial estimate suggests that Respondent 5's trait level is almost 1.4 standard deviations above the mean.

The initial estimates of item difficulties also can be seen as a two-step process. First, we determine the proportion of correct responses for each item. For an item, the proportion of correct responses is the number of respondents who answered the item correctly divided by the total number of respondents who answered the item. For example, Item 1 was answered correctly by five of the six respondents, so Item 1's proportion of correct responses is 5/6 = .83. Table 14.4 presents the proportion of correct responses for each item. To obtain initial estimates of item difficulty, we compute the natural log of the ratio of the proportion of incorrect responses to the proportion of correct responses:

$$\beta_i = \text{In}\left(\frac{1-P_i}{P_i}\right),$$

where P_i is the proportion of correct responses for item i. This initial estimate suggests that Item 1 has a relatively low difficulty level:

$$\beta_i = \text{In}\left(\frac{1-.83}{.83}\right) = \text{In}(.20) = -1.61.$$

This value suggests that even a person with a relatively low level of mathematical ability (i.e., a trait level that is more than 1.5 standard deviations below the mean) will have a 50:50 chance of answering the item correctly. Table 14.4 presents initial estimates of the difficulty level for each of the five items.

As noted, Table 14.4 provides *initial* estimates of ability levels and item difficulties. As also noted earlier, these initial estimates are revised through an iterative series of steps that, ideally, eventually produce final estimates that are as accurate as possible. There are several methods for conducting these series of steps, but they all share a basic framework, as shown in Figure 14.2.

FIGURE 14.2 ⬢ IRT Example: Initial Item Difficulty Estimates and Person Trait-Level Estimates

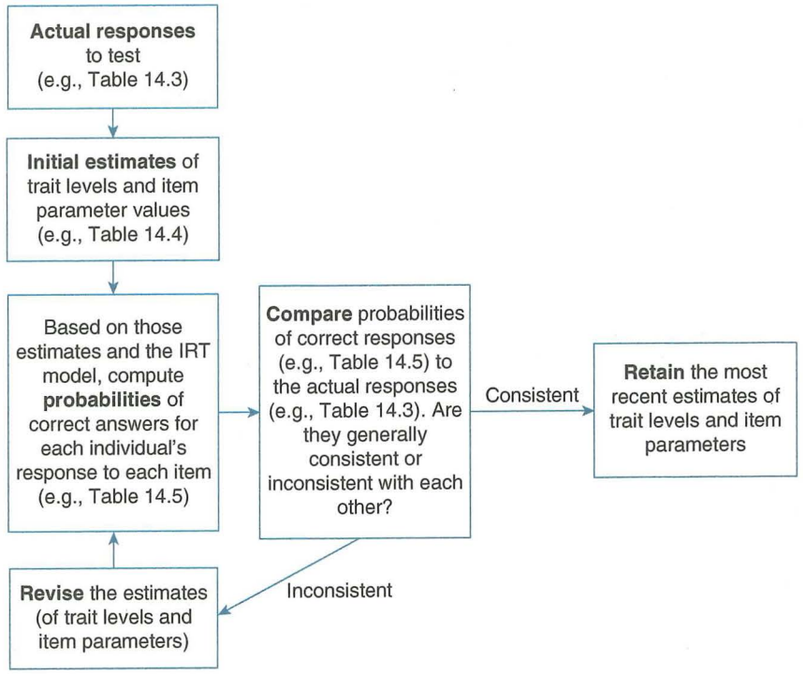


TABLE 14.4 ● IRT Example: Initial Item Difficulty Estimates and Person Trait-Level Estimates

Person	Item 1	Item 2	Item 3	Item 4	Item 5	Proportion Correct	Trait Level
1	1	0	0	0	0	.20	−1.39
2	1	1	0	1	0	.60	.41
3	1	1	1	0	0	.60	.41
4	1	1	0	1	0	.60	.41
5	1	1	1	0	1	.80	1.39
6	0	0	1	0	0	.20	−1.39
Proportion correct	.83	.67	.50	.33	.17		
Difficulty	−1.61	−.69	.00	.69	1.61		

Note: IRT = item response theory.

In this general framework, each step of revision proceeds through a set of substeps. The first substep is to use the initial estimates (of trait levels and item difficulty parameters) to calculate probabilities. That is, the statistical software would compute the probabilities of correct answers for each individual's response to each item. Consider the estimated probability that Individual 1 (with an initial estimated trait level of $\theta = -1.39$) would correctly respond to Item 1 (with an initial estimated difficulty level of −1.61). Entering these initial estimates into the 1PL model (Equation 14.1), we obtain a probability of .56:

$$P(X_{is} = 1 \mid -1.39, -1.61) = \frac{e^{(-1.39-(-1.61))}}{1+e^{(-1.39-(-1.61))}} = \frac{1.25}{1+1.25} = .56.$$

That is, based on the 1PL model and our initial estimates, we would predict that Individual 1 has (slightly) greater than a 50:50 chance of answering Item 1 correctly. Table 14.5 presents all 30 probabilities—one for each of six individuals' response to each of five items.

The second substep in the revision process is to compare these response probabilities to the actual responses provided by the individuals. Returning to Individual 1's response to Item 1, we see in Table 14.3 that this individual did indeed answer Item 1 correctly. In addition, we see in Table 14.5 that Individual 1's predicted probability of a correct response was greater than 50:50. Thus, 1PL model function suggested that the individual was relatively likely to answer the item correctly, and the individual actually did answer the item correctly. Thus, the 1PL model's result is relatively consistent with the empirical fact that the individual actually did correctly respond to that item. This consistency suggests that the 1PL's initial estimate is reasonably accurate, at least for Individual 1.

However, comparing the entirety of Table 14.3 (all actual responses) to the entirety of Table 14.5 (all predicted probabilities of correct responses), you will see noticeable points of *in*consistency. For example, we see that, based on our initial estimates of trait levels and item difficulties, Individual 6 would have only a 20% chance of correctly answering Item 3. However, in fact, this individual did actually answer that item correctly. Thus, this individual correctly answered an item that we would have expected her to answer incorrectly (based on the initial estimates). Similarly, our initial estimates would lead us to expect that Individual 5 would have a relatively good chance of answering Item 4 correctly (probability in Table 14.5 = .67, greater than 50:50). However, this individual did in fact *not* answer the item correctly. In sum, our initial predictions are consistent with some of the actual responses, but they also seem quite inconsistent with other actual responses.

Next, the initial estimates of trait levels and item difficulties would be revised based on the overall pattern of consistencies and inconsistencies. That is, our statistical software would tweak the estimates in a way that would, ideally, increase the overall pattern of consistencies and reduce the number (and magnitude) of inconsistencies. There are a variety of statistical methods for performing these "tweaks," but all of them continue through an iterative process of tweaking the estimates, recalculating probabilities, identifying points of consistency and inconsistency (between predicted probabilities and actual responses), and retweaking the estimates. This process continues until a prespecified point of conclusion is reached.

Again, this procedure is computationally complex, and a full illustration requires a great deal of time and detail. Fortunately, in "real life," statistical software does this computational work on our behalf. We simply supply a set of responses (e.g., Table 14.3), choose an IRT model (e.g., the 1PL model), let the software churn through the process in Figure 14.2, and (hopefully) end up with a set of item parameter estimates (and trait level estimates).

TABLE 14.5 ● IRT Example: Probability of Responses, Based on Initial Item Difficulty Estimates and Person Trait-Level Estimates (From Table 14.4)

Person	Item 1	Item 2	Item 3	Item 4	Item 5
1	.56	.33	.20	.11	.05
2	.88	.75	.60	.43	.23
3	.88	.75	.60	.43	.23
4	.88	.75	.60	.43	.23
5	.95	.89	.80	.67	.44
6	.56	.33	.20	.11	.05

The purpose of the current section has been to provide a relatively intuitive overview of the conceptual logic behind IRT's estimation of key concepts such as respondent trait level and item characteristics. Interested readers can find a thorough, well-described, and relatively accessible illustration using Microsoft Excel at https://www.rasch.org/moulton.htm.

MODEL FIT

Thus far, our discussion has focused on explaining some of the core measurement models of IRT and the concepts at the heart of these models. As we have discussed, test users can choose one of these models and apply it (through statistical software) to item responses from a large sample of respondents to accomplish several goals. Test users might do this for a variety of goals, such as estimating a test (or trait) score for each respondent or obtaining psychometric information about each item and/or about the test as a whole.

However, before putting too much faith in such information, test users must have confidence that their chosen model is indeed appropriate for the actual responses that have been gathered. An earlier section described the general logic and purpose of each model, in terms of binary items versus polytomous items and in terms of the number of item parameters that are included in the models. However, the appropriateness or "fit" of a model is more complex than that.

A full discussion of model fit is far beyond the scope of this chapter (see Maydeu-Olivares, 2015, for an in-depth treatment). For now, the goal is for readers to appreciate the importance of model fit in an IRT analysis and its general logic. As we will see, the use and interpretation of IRT should be done with regard to this issue. Readers interested in more details can refer to several useful sources (e.g., Chernyshenko et al., 2001; Maydeu-Olivares, 2015).

When considering model fit, the key question is whether the actual responses to a set of items are well represented by a given measurement model. Imagine that Beth responds to a set of items, answering all of the easy items correctly, answering the moderately difficult items *in*correctly, but answering two extremely difficult items *correctly*. From an IRT perspective, this pattern of responses does not make much sense. The fact that Beth was able to correctly answer easy items but not the moderately difficult items would lead us—and an IRT model—to guess that she has a relatively low level of the trait. However, the fact that she correctly answered two very difficult items would suggest that she has a *high* level of the trait! So, it's not clear whether Beth has a low level or high level of the trait. Thus, her pattern of actual responses seems quite incompatible with what we would expect and what would be expected based on an IRT model such as the 1PL or 2PL. Based on those models, people who do not correctly answer moderately difficult items are not likely to correctly answer very difficult items. But that's what Beth did.

Thus, we would say that her response pattern does not fit the IRT model. Perhaps many respondents exhibit a pattern that is similarly at odds with what the 1PL or 2PL would predict. In such cases, the model does not seem to fit the sample's responses very well in general.

As described in the previous section and in Figure 14.2, statistical software attempts to find estimates that maximize the consistencies between the predicted pattern of responses and the actual pattern of responses. Indeed, the iterative process unfolds by estimating and reestimating trait levels and item parameters that are as accurate as possible.

Thus, in running an IRT analysis, our statistical software will indeed calculate an estimate of each individual's trait level (and item difficulty levels), but how much should we really trust those estimates—even the final ones that are produced? Our software, say based on using the 1PL, would provide an estimate of Beth's trait level—maybe somewhere between low and high. But is that really warranted, considering her pattern of responses? If the actual responses are incompatible with the 1PL, then perhaps we should not use the 1PL to estimate trait levels based on these responses.

With such issues in mind, test users must obtain, examine, and consider information regarding the "fit" between a particular IRT model and the actual item responses that they are analyzing. There are various ways of evaluating model fit in an IRT-based analysis, and test users should consider these before paying attention to the other information provided by those analyses. Some of the same indices that were noted in Chapter 12 (for CFA) are applied also in an IRT context—chi-square, root mean square error of approximation (RMSEA), and so on. However, fewer such indices are available for IRT (Chernyshenko et al., 2001), and not all such indices are available from all IRT-oriented statistical software.

If the "fit indices" suggest that the model is incompatible with the actual responses, then test users should be very cautious about interpreting other information that is derived from that model (e.g., individuals' trait estimates, items' difficulty estimates, etc.). It is possible that such information is not psychologically meaningful. That said, it is possible that poorly fitting models might provide information that's sufficiently valid for some purposes but not others (Maydeu-Olivares, 2015), and there is some ambiguity about the severity of problems created by poorly fitting models (Sinharay & Haberman, 2014).

If, however, the fit indices suggest good fit between the model and the actual responses, then test users should proceed with interpreting the remaining information as being psychologically meaningful and psychometrically useful. In this way, the issue of model fit in IRT is highly akin to model fit as related to confirmatory factor analysis (see Chapter 12). Upon obtaining evidence of "good fit," test users

might proceed to examine item parameters (difficulty and discrimination), item characteristic curves, item information curves, test information curves, and so on. We discuss many of these in the next section.

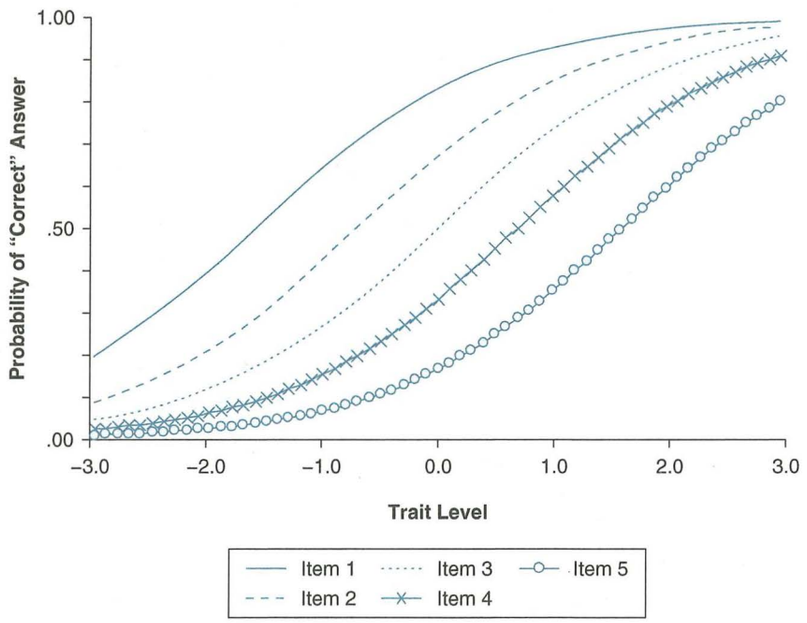
ITEM AND TEST INFORMATION

Returning to the big picture, as a psychometric approach, IRT provides information about items and about tests. In an IRT analysis, item characteristics are combined to reflect characteristics of the test as a whole. In this way, item characteristics such as difficulty and discrimination can be used to evaluate the items and to maximize the overall quality of a test.

Item Characteristic Curves

Psychometricians who use IRT often examine **item characteristic curves (ICCs)** to present and evaluate characteristics of the items on a test. Item characteristic curves, such as those presented in Figure 14.3, reflect the probabilities with which individuals across a range of trait levels are likely to answer each item correctly. The item characteristic curves in Figure 14.3 are based on a 1PL analysis of the five items from the hypothetical mathematics test analyzed earlier (Table 14.3). For item characteristic curves, the horizontal *x*-axis reflects a wide range of trait levels and the vertical *y*-axis reflects probabilities ranging from 0 to 1.0.

FIGURE 14.3 ⬢ Item Characteristic Curves



Each item has a curve, and we can examine an item's curve to find the likelihood that a person with any particular trait level will answer the item correctly. Take a moment to study the curve for Item 1. What is the probability that a person with an average level of mathematical ability will answer the item correctly? To answer this question, we find the point on the Item 1 curve that is directly above the "0" point on the horizontal *x*-axis (recall that the trait level is in *z* score units, so zero is the average trait level), and we see that this point lies between .80 and .90 on the vertical *y*-axis. Thus, a person with an average level of mathematical ability has about an 85% chance of answering Item 1 correctly. Looking at the other curves, we see that a person with an average level of mathematical ability has about a .65 probability of answering Item 2 correctly, a chance of answering Item 3 correctly, and only a .17 probability of answering Item 5 correctly.

Thus, the item characteristic curves provide clues about the likelihoods with which individuals of any trait level would answer any item correctly. Note that the order of the curves, from left to right on the *x*-axis, reflects their difficulty levels. In Figure 14.3, Item 1, with the leftmost curve, is the easiest item, and Item 5, with the rightmost curve, is the most difficult item.

The item characteristic curves are drawn based on the mathematical models, as presented above. For example, the curves in Figure 14.3 are based on the equation for the 1PL or Rasch model (Equation 14.1). To draw an item characteristic curve for an item, we can repeatedly use an IRT response function to compute the probabilities of correct responses for many trait levels. That is, by entering an item's difficulty and a particular trait level (say, –3.0) into the model, we obtain the probability with which a person with that particular trait level will answer that item correctly. We can then enter a different trait level into the model (say, –2.9) and obtain the probability with which a person with the different trait level will answer the item correctly. After conducting this procedure for many different trait levels, we simply plot the probabilities that we have obtained. The line connecting these probabilities reflects the item's characteristic curve. This procedure can be applied for each of the items on the hypothetical five-item test in Table 14.3. Figure 14.3 is based on using the spreadsheet software package Microsoft Excel to compute 305 probabilities for the five items (61 probabilities for each item, each probability representing a different trait level) and to plot the points onto curves.

As noted earlier, the curves presented in Figure 14.3 are based on the simplest model—the 1PL model. This model includes only one item parameter (i.e., difficulty), and it is also designed for items that have only two response possibilities (i.e., binary items). As discussed earlier, other models are designed for more complex measurement scenarios.

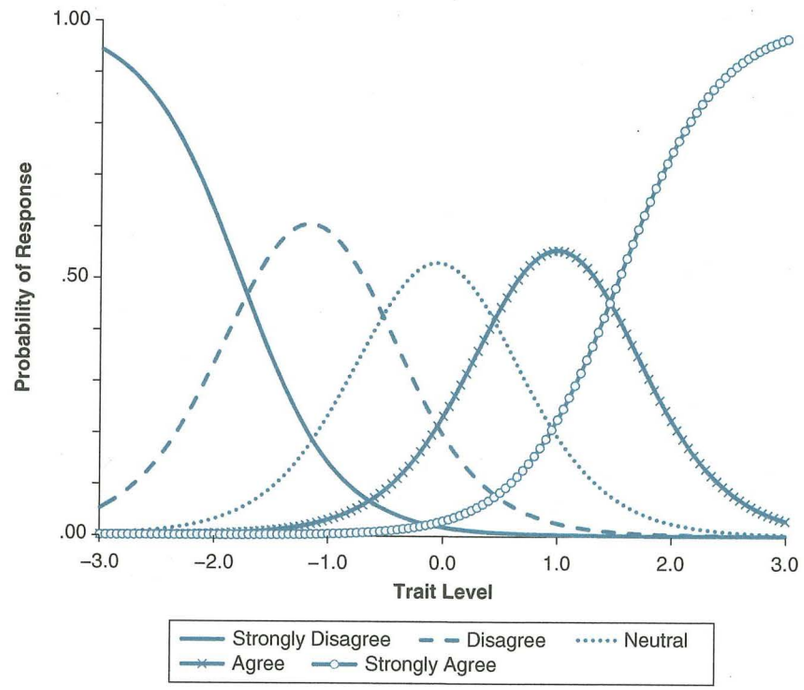
For example, recall that the GRM was designed for items with more than two response options, and it includes difficulty *and* discrimination parameters. Analyses based on this model would produce more complex characteristic

curves for each of a test's items. That is, each item would have several curves—one for each of its response options. Each curve reflects the probabilities with which people of any trait level would choose a particular response option for a particular item.

To illustrate this, an item characteristic curve was plotted for the item that was used earlier to illustrate the GRM ("I enjoy having conversations with friends"). Recall that it had five response options and the following parameter estimates: $\beta_1 = -1.78$, $\beta_2 = -.57$, $\beta_3 = .45$, $\beta_4 = 1.53$, and $\alpha = 2.32$. Based on these values and the GRM, response probabilities were computed for a wide range of trait levels. This was done for each of the item's five response options. Figure 14.4 presents the curves for this item.

As shown in Figure 14.4, each response option has its own curve, representing the probability that people of any particular trait level will endorse that option for the item. Note that as you move from one curve to another, from left to right in the figure, the options increase in their level of endorsement. That is, *strongly disagree* is farthest left, while *strongly agree* is farthest right. This ordering of the curves parallels, reflects, and is driven by the four difficulty parameters (β values) produced by the analysis.

FIGURE 14.4 ● Item Characteristic Curves for an Item, Based on the Graded Response Model



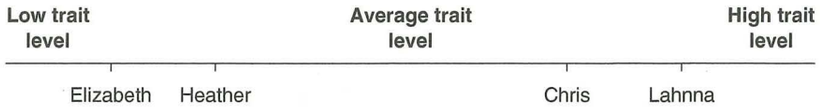
Based on the GRM, each item would have its own *set* of response curves like those in Figure 14.4. These item characteristic curves can convey visual information about the response tendencies associated with each item. Although test developers and users are interested in such tendencies, they are usually even more interested in the psychometric quality of each item and of the test as a whole. To examine this, they evaluate properties called item information and test information, which are closely related to reliability.

Item Information and Test Information

From the perspective of CTT, reliability is an important psychometric consideration for a test. Recall from Chapters 5 and 6 that from the perspective of CTT, we can articulate and estimate the reliability of a test. For example, we might compute coefficient alpha as an estimate of a test's reliability. Importantly, we would compute only *one* reliability estimate for a test, and that one estimate would indicate the degree to which observed test scores are correlated with true scores. That is, from a CTT perspective, we would estimate a single value (i.e., alpha) that reflects the reliability of the entire set of test scores, The idea that there is a single reliability for a particular test is an important way in which CTT differs from IRT.

From the perspective of IRT, a test does not have a single "reliability." Instead, a test might have stronger psychometric quality for some people than for others. That is, a test might provide better information at some trait levels than at other trait levels. This is an important difference between IRT and CTT, and it might require a bit of explanation.

Imagine four people who have different trait levels—Elizabeth, Heather, Chris, and Lahnna. We can depict their relative "true" trait levels along a continuum:



In terms of the underlying psychological trait, Elizabeth and Heather are both below the mean, with a relatively small difference between them. In contrast, Chris and Lahnna are at a relatively high trait level, with a relatively small difference between them.

As emphasized repeatedly throughout this book, the purpose of psychological tests is to detect psychological variability. Indeed, the key goal of most psychological testing is to differentiate (i.e., discriminate between) people with relatively high trait levels from people with lower trait levels. From an IRT perspective, a test provides "good information" when it can accurately detect differences between individuals who have different trait levels.

Even a test that has modest psychometric quality should be able to detect large differences in trait levels. For example, referring to the four individuals above,

a test with moderate psychometric quality should be able to differentiate between the two people with below-average trait scores, on one hand, and the two people with above-average trait scores, on the other. That is, Elizabeth and Heather should score lower than Chris and Lahnna, even if the test is only a modestly good one.

However, if we want to reflect much smaller and more subtle differences between test takers, then we would need a test with stronger psychometric properties. For example, if we wanted to detect the difference between Elizabeth and Heather or if we wanted to know whether Chris's trait level differs from Lahnna's, then the test would need to be quite sensitive to subtle differences in trait levels.

Moreover and importantly, an IRT approach allows for the possibility that a test might be better at reflecting differences at some trait levels rather than at other trait levels. For example, the test might be better at reflecting the difference between Chris and Lahnna than between Elizabeth and Heather. That is, the test might provide better information at high trait levels than at low trait levels.

How could a test provide information that differs by trait level? For example, why would a test be able to discriminate between people who have relatively high trait levels but not between people who have relatively low trait levels? Imagine a two-item test of mathematical ability:

1. What is the square root of 10,000?
2. Solve for x in this equation: $56 = 4x^2 + 3y - 14$.

Both items require a relatively high level of mathematical ability, at least compared with some potential items. If Elizabeth and Heather have low levels of mathematical ability (say, they can both add and subtract, although Heather can do this a bit better than Elizabeth), then they will answer *neither* item correctly. Therefore, Elizabeth and Heather will have the same score on the two-item test and the test cannot differentiate between them. In contrast, Chris and Lahnna have higher levels of mathematical ability, and each might answer at least one item correctly. However, because Lahnna's ability level is a bit higher than Chris's, she might answer *both* items correctly, but Chris might answer only one item correctly. Thus, Chris and Lahnna might have different scores. In terms of "number of items answered correctly," here are the four scores:

	Respondent			
	Elizabeth	Heather	Chris	Lahnna
Items answered correctly	0	0	1	2

Considering this pattern of responding, the test accurately reflects some differences but not others. It accurately differentiates the people with high trait levels (Chris and Lahnna, both of whom answered at least one item correctly) from the people with low trait levels (i.e., Elizabeth and Heather, neither of whom answered

any items correctly). In addition, it differentiates between the people with high trait levels (i.e., Chris, who answered only one item correctly, from Lahnna, who answered two items correctly). However, the test does not differentiate between the people with low trait levels. That is, since Elizabeth and Heather received the same score (presumably a "0" for answering no items correctly), the test is insensitive to the psychological fact that Heather truly has a somewhat higher trait level than does Elizabeth.

In sum, if a test's items have characteristics (e.g., item difficulty levels) that are more strongly represented at some trait levels than at others, then the test's psychometric quality might differ by trait levels. For example, the two-item mathematics test has items with only high difficulty levels, and thus it does not provide clear information discriminating among people at low trait levels. It is insensitive to differences at low trait levels.

We can use IRT to pinpoint the psychometric quality of a test across a wide range of trait levels. This can be seen as a two-step process. First, we evaluate the psychometric quality of each item across a range of trait levels. Just as we can compute the *probability* of a correct answer for an item at a wide range of trait levels (as illustrated in item characteristic curves, discussed earlier), we can use those probabilities to compute *information* at the same range of trait levels. For the 1PL (Rasch) model, **item information** can be computed as follows (Embretson & Reise, 2000):

$$I(\theta) = P_i(\theta)(1 - P_i(\theta)),$$

where $I(\theta)$ is the item's information value at a particular trait level (θ), and $P_i(\theta)$ is the probability that a respondent with a particular trait level will answer the item correctly. For example, Item 1 in Table 14.4 has an estimated difficulty level of −1.61. A person with a trait level that is, say, 3 standard deviations below the mean has a probability of .20 of answering Item 1 correctly (see the equation for computing the probabilities for a Rasch model). Thus, for a trait level of 3 standard deviations below the mean ($\theta = -3$), Item 1 has an information value of .16:

$$I(-3) = .20(1 - .20),$$

$$I(-3) = .16.$$

In contrast, Item 1 has an information value of only .01 at a trait level of 3 standard deviations *above* the mean ($\theta = 3$).

Higher information values indicate greater psychometric quality. Therefore, Item 1 has better psychometric quality at relatively low trait levels than at relatively high trait levels. That is, it is more capable of discriminating among people with low trait levels than among people with high trait levels (presumably because almost everyone with high trait levels will answer the item correctly). Table 14.6 includes probability values and information values that have been computed for each item

at seven trait levels. If we compute information values at many more trait levels, we could display the results in a graph called an **item information curve (IIC)**.

Figure 14.5a presents item information curves for each item in our hypothetical five-item test of mathematics. There are several important issues to consider when interpreting these figures. First, the height of the curve indicates the amount of information that the item provides—higher curves indicate greater psychometric quality. Second, the highest point on a curve represents the trait level at which the item provides the most information. In fact, an item provides the most information at a trait level that corresponds with its difficulty level, estimated earlier. For example, Item 1 (the easiest item) provides the best information at a trait level of –1.61, which is its difficulty level. However, that item does provides very little information at trait levels that are above average. Third, and relatedly, the items often differ in the points at which they provide good information. For example, Item 1 provides good information at relatively low trait levels, Item 3 provides good information at average trait levels, and Item 5 provides good information at relatively high trait levels.

Of course, when we actually use a psychological test, we are concerned with the quality of the test as a whole more than the qualities of individual items. Therefore, we can combine item information values to obtain **test information** values. Specifically, item information values at a particular trait level can be added together to obtain a test information value at that trait level. Again, Table 14.6 provides test information values for our five-item hypothetical test of mathematical ability at seven trait levels, and it shows, for example, that the test information score at an average trait level ($\theta = 0$) is simply the sum of the item information values at this trait level:

$$.97 = .14 + .22 + .25 + .22 + .14.$$

TABLE 14.6 ⬢ IRT Example: Probability of Correct Item Responses, Item Information, and Test Information for Various Trait Levels

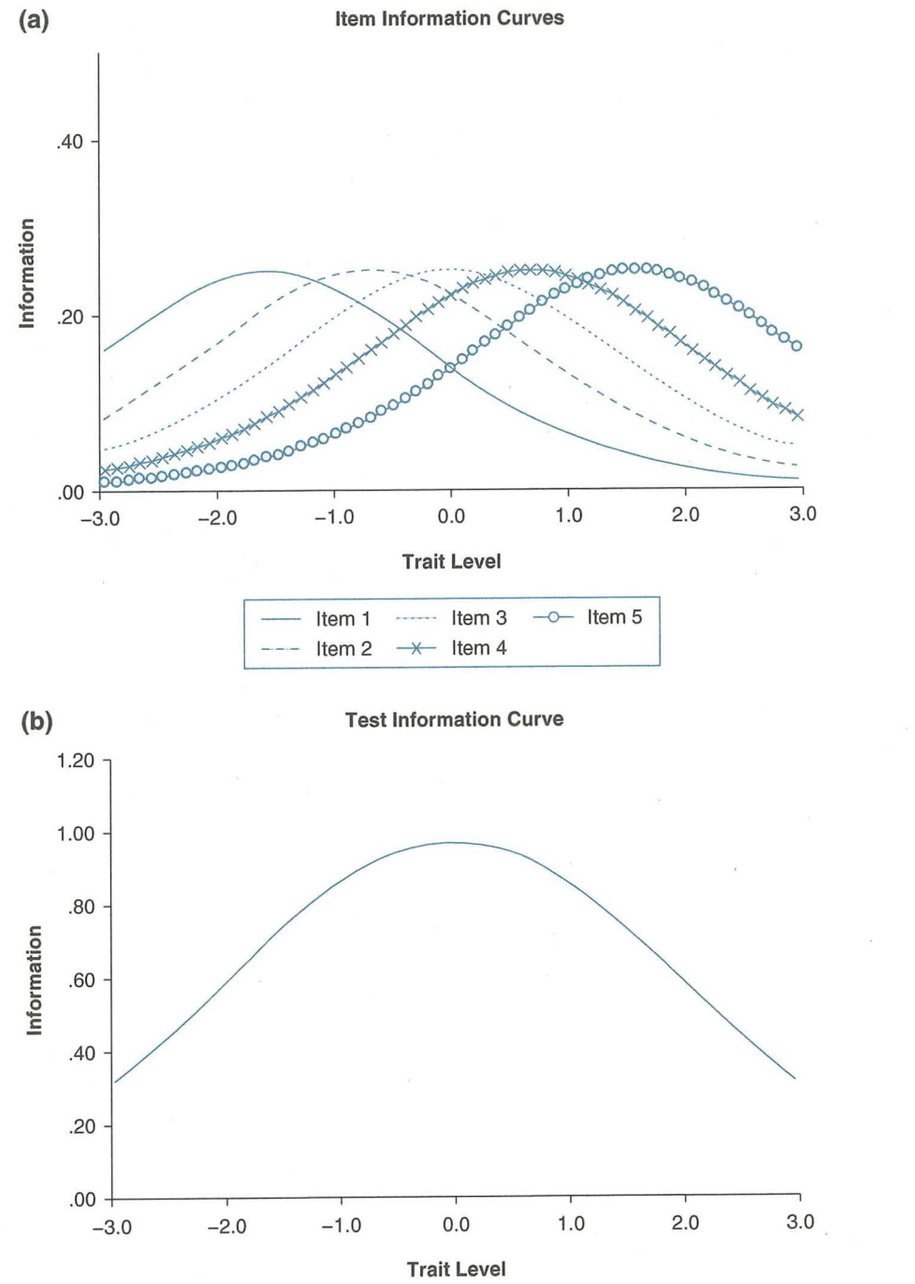
$P(X = 1\|\theta)$ Probability of Correct Answer						Information					
Trait Level	Item 1	Item 2	Item 3	Item 4	Item 5	Item 1	Item 2	Item 3	Item 4	Item 5	Test
–3	.20	.09	.05	.02	.01	.16	.08	.05	.02	.01	.32
–2	.40	.21	.12	.06	.03	.24	.17	.10	.06	.03	.60
–1	.65	.42	.27	.16	.07	.23	.24	.20	.13	.06	.86
0	.83	.67	.50	.33	.17	.14	.22	.25	.22	.14	.97
1	.93	.84	.73	.58	.35	.06	.13	.20	.24	.23	.86
2	.97	.94	.88	.79	.60	.03	.06	.10	.17	.24	.60
3	.99	.98	.95	.91	.80	.01	.02	.05	.08	.16	.32

Note: IRT = item response theory.

Again, if we compute test information scores at many trait levels, we can plot the results in a test information curve, as shown in Figure 14.5b.

A **test information curve (TIC)** is useful for illustrating the degree to which a test provides a different quality of information at different trait levels. Note that

FIGURE 14.5 ⬢ Test and Item Information Curves



our hypothetical test provides the greatest information at an average trait level, and it provides less information at more extreme trait levels. That is, our test does well at differentiating among people who have trait levels within 1 or 2 standard deviations of the mean. In contrast, it is relatively poor at differentiating among people who have trait levels that are more than 2 standard deviations below or above the mean. As noted earlier, Figure 14.5a presents item information curves based on the simplest IRT model—the 1PL or Rasch model. Again, this model includes only one type of item parameter—item difficulty. Thus, we see that the items can provide maximal information at different trait levels. That is, the items differ in their difficulty, and each one has its best psychometric quality at its specific difficulty level.

However, as we have also described, some IRT models include more than one item parameter. For example, the 2PL and the GRM include both difficulty and discrimination parameters for each item. Of course, item and test information can be estimated based on these models as well.

Although the logic behind item information is the same for these more complex models, its computation is more complex. Therefore, we will not delve into the complexities of those computations, but we will consider item information curves that are based on those models (for details on obtaining item information based on the 2PL model, interested readers can see Fraley et al., 2000). Again, this can help expand your familiarity with IRT as a flexible psychometric approach.

For models that include both difficulty and discrimination as item parameters, item information curves can vary in two ways. First, as we have seen for the 1PL model, the curves can vary in terms of location. That is, they can vary in the trait levels at which their maximal information occurs, as reflected in the fact that the curves "top out" at different points along the *x*-axis. Second, item information curves can also vary in terms of height. That is, they can vary in the amount of information that they provide, as reflected in the fact that the curves "top out" at different points along the *y*-axis.

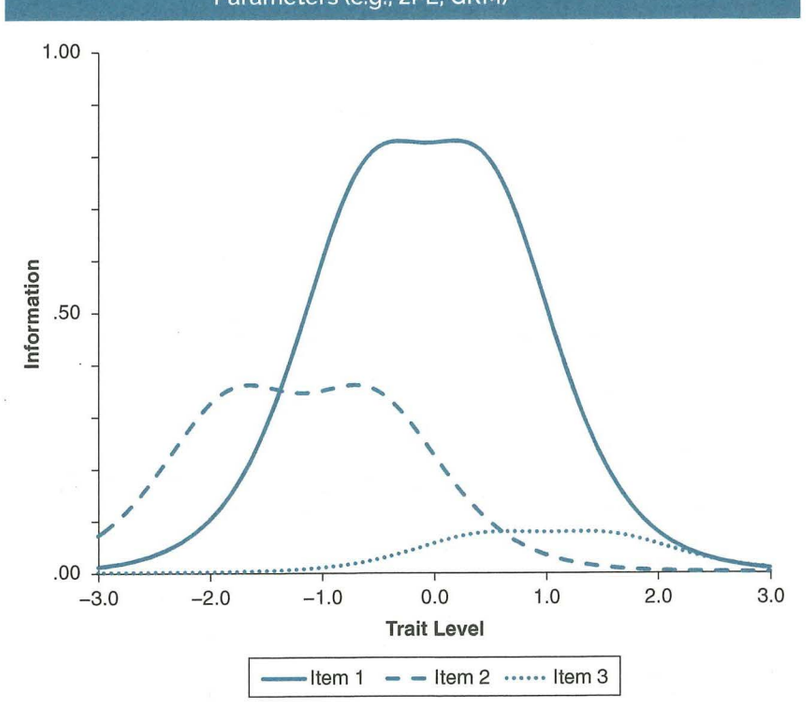
For example, Figure 14.6 presents three item information curves that could be obtained from models such as the 2PL or the GRM. Note that the curves differ in terms of both location and height. From this figure, we see that Item 2 is the easiest item (with its greatest information occurring at approximately 1 standard deviation below the average trait level) and Item 3 is the most difficult (with its greatest information obtained at about 1 standard deviation above the average trait level). We also see that Item 1 provides the best psychometric quality by far. That is, its curve reaches much higher on the information axis than do the other two items. In fact, Item 3 provides very little information, even at its maximal level.

Figures such as Figure 14.6 can provide useful psychometric information for people interested in test development and evaluation. Such figures can highlight

items that might merit revision or deletion from a test. For example, Figure 14.6 suggests that Item 3 provides almost no useful information at any trait level. Based on this finding, we might consider dropping that item or at least examining it in more detail to determine why it performs so poorly (e.g., poor wording, inclusion of construct-irrelevant variance, etc.). In addition, if we decide to drop Item 3, we might want to replace it with an item (or items) that provides good information at high trait levels. Although Item 1 provides good information across a range of trait levels, we might want to be sure to have items that provide good information for making discriminations among people who are 1 to 3 standard deviations above the average trait level. Thus, item information curves can provide very useful insights into items, tests, and test development possibilities.

In sum, take a moment to consider again the difference between IRT and CTT with regard to test reliability. From a CTT perspective, a test has one reliability that can be estimated using an index such as coefficient alpha. From an IRT perspective, a test's psychometric quality can vary across trait levels. This is an important but perhaps underappreciated difference between the two approaches to test theory.

FIGURE 14.6 ⬢ Example of Item Information Curves, Based on a Measurement Model That Includes Discrimination Parameters (e.g., 2PL, GRM)



Note: 2PL = two-parameter logistic model; GRM = graded response model.

APPLICATIONS OF IRT

IRT is a theoretical perspective with tools that have many applications for measurement in a variety of psychological domains. The discussion of item difficulty and discrimination is perhaps most intuitively applied to the measurement of abilities. Indeed, the College Board has used IRT as the basis of the SAT for several years. In addition, within the United States, several states use IRT as the basis of their achievement testing in public school systems. Beyond its application to ability testing, IRT has been applied to domains such as the measurement of attitudes (e.g., Strong et al., 2004) and personality characteristics (Chernyshenko et al., 2001; Fraley et al., 2000; Kashdan et al., 2009).

Test Development and Improvement

A fundamental application of IRT is the evaluation and improvement of the basic psychometric properties of items and tests. Using information about item properties, test developers can select items that reflect an appropriate range of trait levels and that have a strong degree of discriminative ability. Guided by IRT analyses, these selections can create a test with strong psychometric properties across a range of trait levels.

For example, Fraley et al. (2000) used IRT to examine the psychometric properties of four inventories (with a total of 12 subscales) associated with adult attachment. By computing and plotting test information curves for each subscale, Fraley and his colleagues revealed that one inventory in particular, the Experiences in Close Relationships scales (ECR; Brennan et al., 1998), provides a higher level of information than the other inventories. Even further, Fraley and his colleagues used IRT to guide and evaluate modifications to the ECR scales. These modifications produced revised ECR scales with better overall test information quality than the original ECR scales. Notably, this increase in test information was obtained without increasing the number of items.

Differential Item Functioning

In Chapter 11, we discussed test bias. From an IRT perspective, analyses can be conducted to evaluate the presence and nature of differential item functioning (DIF). Differential item functioning occurs when an item's properties in one group are different from the item's properties in another group. For example, DIF exists when a particular item has one difficulty level for males and a different difficulty level for females. Put another way, the presence of DIF means that a male and a female who have the same trait level have different probabilities of answering the item correctly. The existence of DIF between groups indicates that the groups cannot be meaningfully compared on the item.

For example, Smith and Reise (1998) used IRT to examine the presence and nature of DIF for males and females on the Stress Reaction scale of the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982). The Stress Reaction scale assesses the tendency to experience negative emotions such as guilt and anxiety,

and previous research has shown that males and females often have different means on such scales. Smith and Reise argued that this difference could reflect a true gender difference in such traits or that it could be produced by DIF on such scales. Their analysis indicated that although females do appear to have higher trait levels of stress reaction, DIF exists for several items. Furthermore, their analyses revealed interesting psychological meaning for the items that did show DIF. Smith and Reise state that items related to "emotional vulnerability and sensitivity in situations that involve self-evaluation" were easier for females to endorse, but items related to "the general experience of nervous tensions, unexplainable moodiness, irritation, frustration, and being on-edge" (p. 1359) were easier for males to endorse. Smith and Reise concluded that inventories designed to measure negative emotionality will show a large gender difference when "female DIF-type items" are overrepresented and that such inventories will show a small gender difference when "male DIF-type items" are overrepresented. Such insights can inform the development and interpretation of important psychological measures.

Person Fit

Another interesting application of IRT is a phenomenon called **person fit** (Ferrando, 2015; Meijer & Sijtsma, 2001). When we administer a psychological test, we might find a person whose pattern of responses seems strange compared with typical responses. Consider two items that might be found on a measure of friendliness:

1. I like my friends.
2. I am willing to lend my friends as much money as they might ever want.

Most people would probably agree with the first statement (i.e., it is an "easy" item). In contrast, fewer people might agree with the second statement. Although most of us like our friends and would be willing to help them, not all of us would be truly willing to lend our friends "as much money as they might ever want." Certainly, those of us who would lend any amount of money to our friends also would be very likely to state that we like our friends (i.e., endorse the first item). That is, it would not be very strange to find someone who is willing to lend any amount of money to her friends if she also likes her friends, but it would be quite odd to find someone who would be willing to lend any amount of money to her friends if she does not like her friends. There are four possible response patterns for this pair of items, and three of these patterns would have a fairly straightforward interpretation.

Pattern	Item 1	Item 2	Interpretation
1	Disagree	Disagree	Unfriendly person
2	Agree	Disagree	Moderately friendly person
3	Agree	Agree	Very friendly person
4	Disagree	Agree	Unclear interpretation

The analysis of person fit is an attempt to identify individuals whose response pattern does not seem to fit any of the expected patterns of responses to a set of items. Although there are several approaches to the analysis of person fit (Ferrando, 2015; Meijer & Sijtsma, 2001), the general idea is that IRT can be used to estimate item characteristics and then to identify individuals whose responses to items do not adhere to those parameters. For example, IRT analysis might show that Item 1 above has low difficulty (i.e., it does not require a very high level of friendliness to be endorsed) and that Item 2 has higher difficulty. It would be odd to find a person who endorses a difficult item but who does not endorse an easy item.

The identification of individuals with poor person fit to a set of items has several possible implications. Poor person fit could indicate cheating, random responding, low motivation, cultural bias of the test, intentional misrepresentation, or even scoring or administration errors (Schmitt et al., 1999). Furthermore, in a personality assessment context, poor person fit might reveal that a person's personality is unique in that it produces responses that do not fit the "typically expected" pattern of responses (Reise & Waller, 1993).

Computerized Adaptive Testing

An additional application that is commonly associated with IRT is called **computerized adaptive testing (CAT)** (Davey & Pitoniak, 2006; Wainer et al., 2000). CAT is a method of computerized test administration that is intended to provide an accurate and very efficient assessment of individuals' trait levels. Computerized adaptive testing works by using a very large item pool for which IRT has been used to obtain information about the psychometric properties of the items. For example, test administrators might assemble a pool of 300 items and conduct research to estimate the difficulty level for each item. Recall that item difficulty is linked to trait level—an item's difficulty level is the trait level that is required for a respondent to have a .50 probability of answering the item correctly. The information about item difficulties is stored in a computerized database.

As a person begins the test, the computer presents items with difficulty levels targeted at an average trait level (i.e., difficulty levels near zero). From this point, the computer adapts the test to match the individual's apparent trait level. If the individual starts the test with several correct answers, then the computer searches its database of items and selects items with difficulty levels that are a bit above average. These relatively difficult items are then presented to the individual. In contrast, if the individual starts the test with several incorrect answers, then the computer searches its database of items and selects items with difficulty levels that are a bit below average. These relatively easy items are then presented to the individual. Note that the two individuals might respond to two sets of items that are almost completely different.

As the individual continues taking the test, the computer continues to select items that pinpoint the individual's trait level. The computer tracks the individual's responses to specific items with known difficulty levels. By tracking this information,

the computer continually reestimates the individual's trait level as the individual answers some items correctly and others incorrectly. The computer ends the test when it has presented enough items to provide a solid final estimation of the individual's trait level.

Interestingly, the accuracy and efficiency of computerized adaptive tests is obtained by giving different tests to different individuals. This might at first seem counterintuitive, but consider the purpose of adaptive testing. The purpose of adaptive testing is to present items that target each individual's specific trait level efficiently. That is, it presents only the items that really help estimate precisely each examinee's trait level. If a person clearly has a high level of ability, then it is unnecessary to require the individual to respond to very easy questions. Similarly, if a person clearly has a lower level of ability, then we learn nothing by requiring the individual to respond to difficult items. Therefore, instead of presenting one 300-item test to every individual, a CAT program presents each individual with only as many items as are required to pinpoint his or her trait level—probably many fewer than 300 items. Ideally, this method of test administration is more efficient and potentially less aversive for respondents.

Computerized adaptive testing has been used mainly in ability, knowledge, and/or achievement testing. For example, the National Council of State Boards of Nursing (NCSBN) maintains licensure standards for nurses across the United States. Earning a license involves a testing process that uses a pool of nearly 2,000 items with known difficulty levels and a CAT administration process to present items and score respondents. The website for the NCSBN assures candidates for licensure that, as compared with a traditional pencil-and-paper test format, "CAT is able to produce exam results that are more stable using fewer items by targeting items to the candidate's ability" (NCSBN, 2017, p. 14). Similarly, the Graduate Management Admission Test (GMAT) was taken by nearly 250,000 applicants to business schools worldwide in "testing year" 2018 alone (Graduate Management Admission Council [GMAC], 2019), and as of this writing, it is primarily administered through computerized adaptive testing. To explain the adaptive nature of the test, the website for the GMAT informs test takers that "the GMAT exam is computer adaptive, which means it selects each question for the test taker based on his or her ability level. This makes the GMAT Total score an extremely precise measure of an individual's ability. It's far more efficient than a paper test, on which everyone answers the same questions" (GMAC, n.d.). Clearly, CAT is a practically useful application of IRT, and its use seems to be growing.

TECHNICAL APPENDIX: R SYNTAX

This chapter presented item response theory (IRT) as a modern perspective on psychometric quality. The following syntax demonstrates how R can be used to conduct IRT-related analyses. It demonstrates this in some detail for the 1PL/Rasch model, estimating item parameters and individuals' trait levels, evaluating model fit, and printing several plots.