Statistics for Social and Behavioral Sciences

Advisors:

S.E. Fienberg
W.J. van der Linden

For other titles published in this series, go to
http://www.springer.com/3463

Mark D. Reckase

Multidimensional Item Response Theory

Springer

Mark D. Reckase
Michigan State University
Counseling, Educational, Psychology,
and Special Education Department
461 Erickson Hall
East Lansing MI 48824-1034
USA

ISBN 978-0-387-89975-6 e-ISBN 978-0-387-89976-3
DOI 10.1007/978-0-387-89976-3
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009927904

Preface

Item response theory (IRT) is a general framework for specifying mathematical functions that describe the interactions of persons and test items. It has a long history, but its popularity is generally attributed to the work of Fredrick Lord and Georg Rasch starting in the 1950s and 1960s. Multidimensional item response theory (MIRT) is a special case of IRT that is built on the premise that the mathematical function includes as parameters a vector of multiple person characteristics that describe the skills and knowledge that the person brings to a test and a vector of item characteristics that describes the difficulty of the test item and the sensitivity of the test item to differences in the characteristics of the persons. MIRT also has a long history, going back to the work of Darrel Bock, Paul Horst, Roderick McDonald, Bengt Muthen, Fumiko Samajima, and others starting in the 1970s.

The goal of this book is to draw together in one place the developments in the area of MIRT that have occurred up until 2008. Of course, it is not possible to be totally comprehensive, but it is believed that most of the major developments have been included.

The book is organized into three major parts. The first three chapters give background information that is useful for the understanding of MIRT. Chapter 1 is a general conceptual overview. Chapter 2 provides a summary of unidimensional IRT. Chapter 3 provides a summary of the historical underpinnings of MIRT. Chapter 2 can be skipped if the reader already has familiarity with IRT. Chapter 3 provides useful background, but it is not required for understanding of the later chapters.

The second part of the book includes Chaps. 4–6. These chapters describe the basic characteristics of MIRT models. Chapter 4 describes the mathematical forms of the models. Chapter 5 summarizes the statistics that are used to describe the way that test items function within an MIRT context. Chapter 6 describes procedures for estimating the parameters for the models.

The third part of the book provides information needed to apply the models and gives some examples of applications. Chapter 7 addresses the number of dimensions needed to describe the interactions between persons and test items. Chapter 8 shows how to define the coordinate system that is used to locate persons in a space relative to the constructs defined by the test items. Chapter 9 describes methods for converting parameter estimates from different MIRT calibrations to the same coordinate system. Finally, Chap. 10 shows how all of these procedures can be applied in the context of computerized adaptive testing.

Chapters 4–9 have been used for a graduate level course in MIRT. In the context of such a course, Chap. 10 can be used as an example of the application of the methodology. The early chapters of the book are a review of basic concepts that advanced graduate students should know, but that need to be refreshed.

Chapters 7–9 should be particularly useful for those who are interested in using MIRT for the analysis and reporting of large-scale assessment results. Those chapters lay out the procedures for specifying a multidimensional coordinate space and for converting results from subsequent calibrations of test forms to that same coordinate system. These are procedures that are needed to maintain a large-scale assessment system over years. The content of these chapters also addresses methods for reporting subscores using MIRT.

There are many individuals that deserve some credit for the existence of this book. First, my wife, Char Reckase, did heroic labors proofing the first drafts of the full manuscript. This is second only to the work she did typing my dissertation back in the days before personal computers. Second, the members of the STAR Department at ACT, Inc. helped with a lot of the early planning of this book. Terry Ackerman, Jim Carlson, Tim Davey, Ric Leucht, Tim Miller, Judy Spray, and Tony Thompson were all part of that group and did work on MIRT. Several of them even agreed to write chapters for an early version of the book – a few even finished first drafts. Although I profited from all of that work, I decided to start over again several years ago because there had been a substantial increase in new research on MIRT.

The third contributors to the book were the graduate students who reacted to first drafts of the chapters as part of my graduate courses in IRT and an advanced seminar in MIRT. Many students contributed and there are too many to list here. However, Young Yee Kim, Adam Wyse, and Raymond Mapuranga provided much more detailed comments than others and need to be honored for that contribution.

East Lansing, MI *M.D. Reckase*

Contents

Chapter 1
Introduction

Test items are complicated things. Even though it is likely that readers of this book will know what test items are from their own experience, it is useful to provide a formal definition.

> "A test item in an examination of mental attributes is a unit of measurement with a stimulus and a prescriptive form for answering; and, it is intended to yield a response from an examinee from which performance in some psychological construct (such as ability, predisposition, or trait) may be inferred."
>
> Osterlind 1990, p. 3

The definition of a test item itself is complex, but it does contain a number of clear parts – stimulus material and a form for answering. Usually, the stimulus material asks a specific question and the form for answering yields a response. For most tests of achievement, aptitude, or other cognitive characteristics, the test item has a correct answer and the response is scored to give an item score.

To show the complexity of a test item and clarify the components of a test item, an example is provided. The following test item is a measure of science achievement and the prescriptive form for answering is the selection of an answer choice from a list. That is, it is a multiple-choice test item.

Which of the following is an example of a chemical reaction?

A. A rainbow
B. Lightning
C. Burning wood
D. Melting snow

Selecting a response alternative for this test item is thought of as the result of the interaction between the capabilities of the person taking the test and the characteristics of the test item. This test item requires different types of knowledge and a number of skills. First, persons interacting with this test item, that is, working to determine the correct answer, must be able to read and comprehend English. They need to understand the question format. They need to know the meaning of "chemical reaction," and the meanings of the words that are response alternatives. They need to understand that they can only make one choice and the means for recording the choice. They need to know that a rainbow is the result of refracting light,

M.D. Reckase, *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-89976-3_1,

lightning is an electrical discharge, melting snow is a change of state for water, but burning wood is a combination of the molecular structure of wood with oxygen from air to yield different compounds. Even this compact science test item is very complex. Many different skills and pieces of knowledge are needed to identify the correct response. This type of test item would typically be scored 1 for selecting the correct response, C, and 0 for selecting any other choice. The intended meaning of the score for the test item is that the person interacting with the item either has enough of all of the necessary skills and knowledge to select the correct answer, or that person is deficient in some critical component. That critical component could be reading skill or vocabulary knowledge, or knowledge of the testing process using multiple-choice items. The author of the item likely expects that the critical component has to do with knowledge of chemical reactions.

Test items are complicated devices, while people are even more complex. The complexities of the brain are not well understood, but different people probably have different "wiring." Their neural pathways are probably not organized in the same way. Further, from their birth, or maybe even from before birth, they have different experiences and learn different things from them. On the one hand, people who have lived all of their lives in a hot climate may have never watched snow melt. On the other hand, those from cold climates may recognize many different types of snow. From all of their experiences, people develop complex interrelationships between the pieces of information they have acquired and the methods they have for retrieving and processing that information. No two people likely have the same knowledge base and use the same thought processes when interacting with the test item presented on the previous page.

A test consists of collections of test items. Each of the test items is complex in its own way. The people who take a test also consist of very diverse individuals. Even identical twins will show some differences in their knowledge and skills because their life experiences are not exactly the same after their birth. The interactions of test takers with the test items on a test result in a set of responses that represent very complex processes.

Early procedures for test analysis were based on very simple methods such as counting the number of correct responses in the scored set of items. The assumption was that people with more correct responses (each counting for one point) had more of a particular ability or skill than those who had fewer correct responses. Answering one test item correctly added the same amount to the number of correct responses as answering any other item correctly.

Those who analyze test data have always known that some test items are more difficult than others. To capture the observed differences in test items, more complex ways of describing test item functioning were developed. Measures of item discriminating power, the proportion of persons choosing each alternative, and a selection of other statistical indicators are regularly collected to describe the function of test items. Item response theory methods have also been developed. These methods describe the functioning of test items for people at different levels on a hypothesized continuum of skill or knowledge. All of these methods provide relatively simple summaries of the complex interaction between complicated people

and complicated test items. It is the purpose of this book to provide methods that describe these interactions in ways that more realistically depict the complexity of the data resulting from the administration of tests to people.

1.1 A Conceptual Framework for Thinking About People and Test Items

People vary in many different ways. The focus in this book will be on measuring the ways people differ in their cognitive skills and knowledge, although many of the methods also apply to measuring attitudes, interests, and personality characteristics as well. It will be left to others to generalize the methods to those other targets of measurement. Although it might be argued that for some skills and knowledge a person either has that particular skill or knowledge or does not, in this book it is assumed that people vary in the degree to which they have a skill or the degree to which they have knowledge. For the item presented on the first page of this chapter, a person may know a lot about chemical reactions or very little, or have varying degrees of knowledge between those extremes. They may have varying levels of English reading comprehension. It will be assumed that large numbers of people can be ordered along a continuum of skill or knowledge for each one of the many ways that people differ.

From a practical perspective, the number of continua that need to be considered depends on the sample of people that is of interest. No continuum can be defined or detected in item response data if people do not vary on that particular skill or knowledge. For example, a group of second grade students will probably not have any formal knowledge of calculus so it will be difficult to define a continuum of calculus knowledge or skill based on an ordering of second grade students on a calculus test. Even though it might be possible to imagine a particular skill or knowledge, if the sample, or even the population, of people being considered does not vary on that skill or knowledge, it will not be possible to identify that continuum based on the responses to test items from that group. This means that the number of continua that need to be considered in any analysis of item response data is dependent on the sample of people who generated those data. This also implies that the locations of people on some continua may have very high variability while the locations on others will not have much variability at all.

The concept of continuum that is being used here is similar to the concept of "hypothetical construct" used in the psychological literature (MacCorquodale and Meehl 1948). That is, a continuum is a scale along which individuals can be ordered. Distances along this continuum are meaningful once an origin for the scale and units of measurement are specified. The continuum is believed to exist, but it is not directly observable. Its existence is inferred from observed data; in this case the responses to test items. The number of continua needed to describe the differences in people is assumed to be finite, but large. In general, the number of continua on which a group of people differ is very large and much larger than could be measured with any actual test.

The number of continua that can be defined from a set of item response data is not only dependent on the way that the sample of test takers vary, but it is also dependent on the characteristics of the test items. For test items to be useful for determining the locations of people in the multidimensional space, they must be constructed to be sensitive to differences in the people. The science item presented on first page of this chapter was written with the intent that persons with little knowledge of chemical reactions would select a wrong response. Those that understood the meaning of the term "chemical reaction" should have a high probability of selecting response C. In this sense, the item is expected to be sensitive to differences in knowledge of chemical reactions. Persons with different locations on the cognitive dimension related to knowledge of chemical reactions should have different probabilities of selecting the correct response.

The test item might also be sensitive to differences on other cognitive skills. Those who differ in English reading comprehension might also have different probabilities of selecting the correct response. Test items may be sensitive to differences of many different types. Test developers expect, however, that the dimensions of sensitivity of test items are related to the purposes of measurement. Test items for tests that have important consequences, high stakes tests, are carefully screened so that test items that might be sensitive to irrelevant differences, such as knowledge of specialized vocabulary, are not selected. For example, if answer choice C on the test item were changed to "silage," students from farm communities might have an advantage because they know that silage is a product of fermentation, a chemical process. Others might have a difficult time selecting the correct answer, even though they knew the concept "chemical reaction."

Ultimately, the continua that can be identified from the responses to test items depend on both the number of dimensions of variability within the sample of persons taking the test and the number of dimensions of sensitivity of the test items. If the test items are carefully crafted to be sensitive to differences in only one cognitive skill or type of knowledge, the item response data will only reflect differences along that dimension. If the sample of people happens to vary along only one dimension, then the item response data will reflect only differences on that dimension. The number of dimensions of variability that are reflected in test data is the lesser of the dimensions of variability of the people and the dimensions of sensitivity of the test items.

The ultimate goal of the methods discussed in this book is to estimate the locations of individuals on the continua. That is, a numerical value is estimated for each person on each continuum of interest that gives the relative location of persons on the continuum. There is some confusion about the use of the term "dimensions" to refer to continua and how dimensions relate to systems of coordinates for locating a person in a multidimensional space. To provide a conceptual framework for these distinctions, concrete examples are used that set aside the problems of defining hypothetical constructs. In later chapters, a more formal mathematical presentation will be provided.

To use a classic example (Harman 1976, p. 22), suppose that very accurate measures of length of forearm and length of lower leg are obtained for a sample of girls

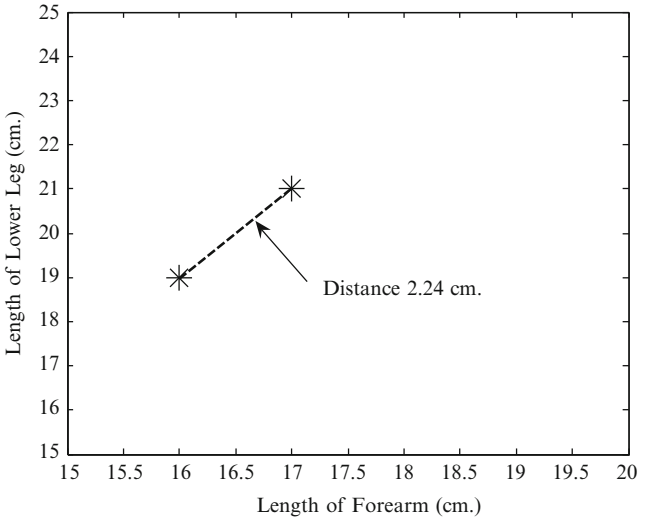

Fig. 1.1 Distance between two girls based on arm and leg lengths

with ages from seven to 17. Certainly, each girl can be placed along a continuum using each of these measures and their ordering along the two continua would not likely be exactly the same. In fact, Harman reports the correlation between these two measures as 0.801. Figure 1.1 shows the locations of two individuals from the sample of 300 girls used for the example. In this case, the lengths of forearm and lower leg can be considered as dimensions of measurement for the girls in this study. The physical measurements are also coordinates for the points in the graph. In general, the term "coordinate" will be considered as numerical values that are used to identify points in a space defined by an orthogonal grid system. The term dimension will be used to refer to the scale along which meaningful measurements are made. Coordinate values might not correspond to measures along a dimension.

For this example, note the obvious fact that the axes of the graph are drawn at right angles (i.e., orthogonal) to each other. The locations of the two girls are represented by plotting the pairs of lengths (16, 19) and (17, 21) as points. Because the lengths of arm and leg are measured in centimeters, the distance between the two girls in this representation is also in centimeters. This distance does not have any intrinsic meaning except that large numbers mean that the girls are quite dissimilar in their measurements and small numbers mean they are more similar on the measures. Height and weight could also be plotted against each other and then the distance measure would have even less intrinsic meaning. The distance measure, D, in this case was computed using the standard distance formula,

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \tag{1.1}$$

where the x_i and y_i values refer to the first and second values in the order pairs of values, respectively, for $i = 1, 2$.

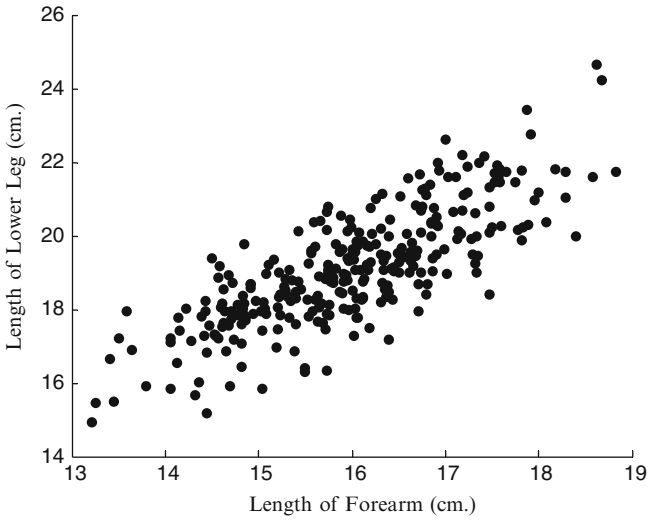

Fig. 1.2 Scatter plot of arm and leg lengths for 300 girls with ages from 7 to 17

The use of the standard distance formula is based on the assumption that the coordinate axes are orthogonal to each other. If that were not the case, the distance formula would need another term under the radical that accounts for the angle between the axes. Although this may seem like a trivial point, it is very important. Coordinate axes are typically made orthogonal to each other so that the standard distance formula applies. However, having orthogonal axes does not mean that the values associated with the coordinate axes (e.g., the coordinates used to plot points) are uncorrelated. In fact, for the data provided by Harman (1976), the correlation between coordinates for the points is 0.801. This correlation is represented in Fig. 1.2.

The correlation between coordinates has nothing to do with the mathematical properties of the frame of reference used to plot them. Using orthogonal coordinate axes means that the standard distance formula can be used to compute the distance between points. The correlations between coordinates in this orthogonal coordinate system can take on any values in the range from -1 to 1. The correlation is a descriptive statistic for the configuration of the points in this Cartesian coordinate space. The correlation does not describe the orientation of the coordinate axes.

The coordinate system does not have to be related to specific continua (e.g., the dimensions) that are being measured. In fact, for the mathematical representations of the continua defined by test results, it will seldom be the case that the constructs being measured exactly line up with the coordinate axes. This is not a problem and using an arbitrary set of coordinate axes is quite common. An example is the system of latitude and longitude that is used to locate points on a map. That system does not have any relationship to the highways or streets that are the ways most people move from place to place or describe the locations of places. The latitude and longitude system is an abstract system that gives a different representation of the observed system of locations based on highways and streets.

Suppose someone is traveling by automobile between two cities in the United States, St. Louis and Chicago. The quickest way to do this is to drive along Interstate Highway 55, a direct route between St. Louis and Chicago. It is now standard that the distances along highways in the United States are marked with signs called mile markers every mile to indicate the distance along that highway. In this case, the mile markers begin at 1 just across the Mississippi River from St. Louis and end at 291 at Chicago. These signs are very useful for specifying exits from the highway or locating automobiles stopped along the highway. In the context here, the mile markers can be thought of as analogous to scores on an achievement test that show the gain in knowledge (the intellectual distance traveled) by a student.

A map of highway Interstate 55 is shown in Fig. 1.3. Note that the highway does not follow a cardinal direction and it is not a straight line. Places along the highway can be specified by mile markers, but they can also be specified by the coordinates of latitude and longitude. These are given as ordered pairs of numbers in parentheses. In Fig. 1.3, the two ways of locating a point along the highway are shown for

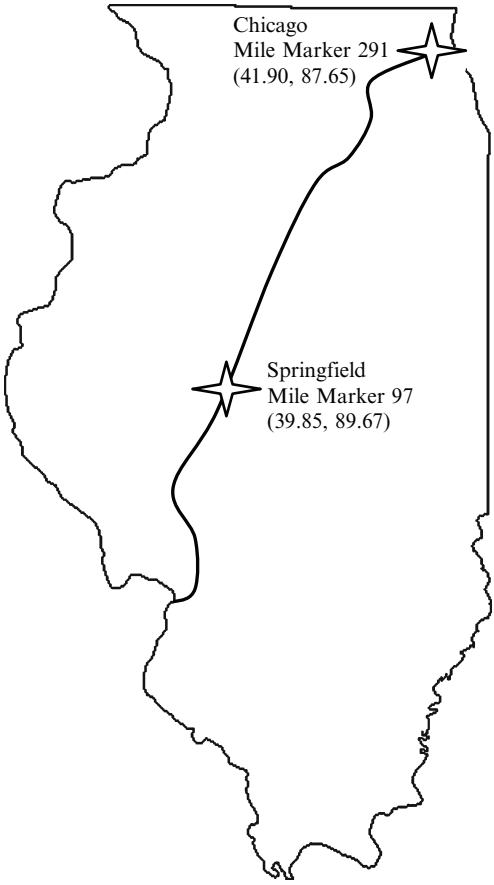

Fig. 1.3 Locations of cities along Interstate Highway 55 from East St. Louis to Chicago

the cities of Springfield, Illinois and Chicago, Illinois, USA. Locations can be specified by a single number, the nearest mile marker, or two numbers, the measures of latitude and longitude.

It is always possible to locate points using more coordinates than is absolutely necessary. We could add a third coordinate giving the distance from the center of the Earth along with the latitude and longitude. In the state of Illinois, this coordinate would have very little variation because the ground is very flat, and it is probably irrelevant to persons traveling from St. Louis to Chicago. But it does not cause any harm either. Using too few coordinates can cause problems, however. The coordinates of my New York City apartment where I am writing this chapter while on sabbatical are 51st Street, 7th Avenue, and the 14th floor, (51, 7, 14). If only (7,14) is listed when packages are to be delivered, they do not uniquely identify me. It is unlikely that they will ever arrive. Actually, there is even a fourth coordinate (51, 7, 14, 21) to identify the specific apartment on the 14th floor. Using too few coordinates results in an ambiguous specification of location.

The purpose of this book is to describe methodology for representing the locations of persons in a hypothetical multidimensional cognitive space. As this methodology is described, it is important to remember that a coordinate system is needed to specify the locations of persons, but it is not necessary to have the minimum number of coordinates to describe the location, and coordinates do not necessarily coincide with meaningful psychological dimensions. The coordinate system will be defined with orthogonal axes, and the Euclidean distance formula will be assumed to hold for determining the distance between points. The coordinates for a sample of persons may have nonzero correlations, even though the axes are orthogonal.

1.2 General Assumptions Behind Model Development

The methodology described in this book defines mathematical functions that are used to relate the location of a person in a multidimensional Cartesian coordinate space to the probability of generating a correct response to a test item. This relationship is mediated by the characteristics of the test item. The characteristics of the test item will be represented by a series of values (parameters) that are estimated from the item response data. The development of the mathematical function is based on a number of assumptions. These assumptions are similar to those presented in most item response theory books, but some additional assumptions have been included that are not typically explicitly stated. This is done to make the context of the mathematical formulation as clear as possible.

The first assumption is that the location of the persons being assessed does not change during the process of taking the test. This assumption may not be totally true in practice – examinees may learn something from interacting with the items that change their locations, or there may be other events that take place in the examination setting (e.g., cheating, information available in the room, etc.) that results in some learning. It may be possible to develop models that can capture changes

during the process of the test, but that is beyond the scope of the models presented here. There are models that consider the changes from one test session to the next (Embretson 1991; Fischer 1995b). These will be placed in the larger context of multidimensional item response theory models.

The second assumption is that the characteristics of a test item remain constant over all of the testing situations where it is used. This does not mean that the observed statistics used to summarize item performance will remain constant. Certainly, the proportion correct for an item will change depending on the capabilities of the sample of examinees. The difficulty of the item has not changed, but the difficulty has a different representation because of differences in the examinee sample. Similarly, a test item written in English may not function very well for students who only comprehend text written in Spanish. This suggests that one of the characteristics of the item is that it is sensitive to differences in language skills of the examinee sample, even if that is not the clear focus of the item. This means that in the full multidimensional representation of the item characteristics, there should be an indicator of the sensitivity to differences in language proficiency. When the examinee sample does not differ on language proficiency, the sensitivity of test item to such differences will not be detectable in the test data. However, when variation exists in the examinee population, the sensitivity of the test item to that variation will affect the probability of correct response to the test item.

A third assumption is that the responses by a person to one test item are independent of their responses to other test items. This assumption is related to the first assumption. Test items are not expected to give information that can improve performance on later items. Similarly, the responses generated by one person are assumed to not influence the responses of another person. One way this could occur is if one examinee copies the responses of another. It is expected that the control of the testing environment is such that copying or other types of collaboration do not occur. The third assumption is labeled "local independence" in the item response theory literature. The concept of local independence will be given a formal definition in Chap. 6 when the procedures for estimating parameters are described.

A fourth assumption is that the relationship between locations in the multidimensional space and the probabilities of correct response to a test item can be represented as a continuous mathematical function. This means that for every location there is one and only one value of probability of correct response associated with it and that probabilities are defined for every location in the multidimensional space – there are no discontinuities. This assumption is important for the mathematical forms of models that can be considered for representing the interaction between persons and test items.

A final assumption is that the probability of correct response to the test item increases, or at least does not decrease, as the locations of examinees increase on any of the coordinate dimensions. This is called the "monotonicity" assumption and it seems reasonable for test items designed for the assessment of cognitive skills and knowledge. Within IRT, there are models that do not require this assumption (e.g., Roberts et al. 2000). The generalization of such models to the multidimensional case is beyond the scope of this book.

The next several chapters of this book describe the scientific foundations for the multidimensional item response theory models and several models that are consistent with the listed assumptions. These are not the only models that can be developed, but also they are models that are currently in use. There is some empirical evidence that these models provide reasonable representations of the relationship between the probability of correct response to a test item and the location of a person in a multidimensional space. If that relationship is a reasonable approximation to reality, practical use can be made of the mathematical models. Such applications will be provided in the latter chapters of the book.

1.3 Exercises

1. Carefully read the following test item and select the correct answer. Develop a list of all of the skills and knowledge that you believe are needed to have a high probability of selecting the correct answer.

The steps listed below provide a recipe for converting temperature measured in degrees Fahrenheit (F) into the equivalent in degrees Celsius (C).

1. Subtract 32 from a temperature given in degrees Fahrenheit.
2. Multiply the resulting difference by 5.
3. Divide the resulting product by 9.

Which formula is a correct representation of the above procedure?

A. $C = F - 32 \times 5/9$
B. $C = (F - 32) \times 5/9$
C. $C = F - (32 \times 5)/9$
D. $C = F - 32 \times (5/9)$
E. $C = F - (32 \times 5/9)$

2. In our complex society, it is common to identify individuals in a number of different ways. Sometimes it requires multiple pieces of information to uniquely identify a person. For example, it is possible to uniquely identify students in our graduate program from the following information: year of entry, gender (0,1), advisor, office number – (2004, 1, 3, 461). Think of ways that you can be uniquely identified from strings of numbers and other ways you can be identified with one number.

3. Which of the following mathematical expressions is an example of a function of x and which is not? Give the reasons for your classification.

A. $y = x^3 - 2x^2 + 1$
B. $y^2 = x$
C. $z^2 = x^2 + y^2$

Chapter 2
Unidimensional Item Response Theory Models

In Chap. 3, the point will be made that multidimensional item response theory (MIRT) is an outgrowth of both factor analysis and unidimensional item response theory (UIRT). Although this is clearly true, the way that MIRT analysis results are interpreted is much more akin to UIRT. This chapter provides a brief introduction to UIRT with a special emphasis on the components that will be generalized when MIRT models are presented in Chap. 4. This chapter is not a thorough description of UIRT models and their applications. Other texts such as Lord (1980), Hambleton and Swaminathan (1985), Hulin et al. (1983), Fischer and Molenaar (1995), and van der Linden and Hambleton (1997) should be consulted for a more thorough development of UIRT models.

There are two purposes for describing UIRT models in this chapter. The first is to present basic concepts about the modeling of the interaction between persons and test items using simple models that allow a simpler explication of the concepts. The second purpose is to identify shortcomings of the UIRT models that motivated the development of more complex models. As with all scientific models of observed phenomena, the models are only useful to the extent that they provide reasonable approximations to real world relationships. Furthermore, the use of more complex models is only justified when they provide increased accuracy or new insights. One of the purposes of this book is to show that the use of the more complex MIRT models is justified because they meet these criteria.

2.1 Unidimensional Models of the Interactions of Persons and Test Items

UIRT comprises a set of models (i.e., item response theories) that have as a basic premise that the interactions of a person with test items can be adequately represented by a mathematical expression containing a single parameter describing the characteristics of the person. The basic representation of a UIRT model is given in (2.1). In this equation, θ represents the single parameter that describes the characteristics of the person, $\boldsymbol{\eta}$ represents a vector of parameters that describe the characteristics of the test item, U represents the score on the test item, and u is

M.D. Reckase, *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-89976-3_2,

a possible value for the score, and f is a function that describes the relationship between the parameters and the probability of the response, P($U = u$).

$$P(U = u \mid \theta) = f(\theta, \boldsymbol{\eta}, u). \tag{2.1}$$

The item score, u, appears on both sides of the equation because it is often used in the function to change the form of the function depending on the value of the score. This is done for mathematical convenience. Specific examples of this use will be provided later in this chapter.

The assumption of a single person parameter for an IRT model is a strong assumption. A substantial amount of research has been devoted to determining whether this assumption is reasonable when modeling a particular set of item response data. One type of research focuses on determining whether or not the data can be well modeled using a UIRT model. For example, the DIMTEST procedure developed by Stout et al. (1999) has the purpose of statistically testing the assumption that the data can be modeled using a function like the one given in (2.1) with a single person parameter. Other procedures are available as well (see Tate 2003 for a summary of these procedures). The second type of research seeks to determine the effect of ignoring the complexities of the data when applying a UIRT model. These are generally robustness studies. Reckase (1979) presented one of the first studies of this type, but there have been many others since that time (e.g., Drasgow and Parsons 1983; Miller and Linn 1988; Yen 1984).

Along with the assumption of a single person parameter, θ, most UIRT models assume that the probability of selecting or producing the correct response to a test item scored as either correct or incorrect increases as θ increases. This assumption is usually called the monotonicity assumption. In addition, examinees are assumed to respond to each test item as an independent event. That is, the response by a person to one item does not influence the response to an item produced by another person. Also, the response by a person to one item does not affect that person's tendencies to respond in a particular way to another item. The response of any person to any test item is assumed to depend *solely* on the person's single parameter, θ, and the item's vector of parameters, $\boldsymbol{\eta}$. The practical implications of these assumptions are that examinees do not share information during the process of responding to the test items, and information from one test item does not help or hinder the chances of correctly responding to another test item. Collectively, the assumption of independent responses to all test items by all examinees is called the local independence assumption.

The term "local" in the local independence assumption is used to indicate that responses are assumed independent at the level of individual persons with the same value of θ, but the assumption does not generalize to the case of variation in θ. For groups of individuals with variation in the trait being assessed, responses to different test items typically are correlated because they are all related to levels of the individuals' traits. If the assumptions of the UIRT model hold, the correlation between item scores will be solely due to variation in the single person parameter.

The implication of the local independence assumption is that the probability of a collection of responses (responses of one person to the items on a test, or the responses of many people to one test item) can be determined by multiplying the probabilities of each of the individual responses. That is, the probability of a vector of item responses, $\mathbf{u}$, for a single individual with trait level θ is the product of the probabilities of the individual responses, u_i, to the items on a test consisting of I items.

$$P(\mathbf{U} = \mathbf{u} \mid \theta) = \prod_{i=1}^{I} P(u_i \mid \theta) = P(u_1 \mid \theta) P(u_2 \mid \theta) \cdots P(u_I \mid \theta), \tag{2.2}$$

where $\mathrm{P}(\mathbf{U} = \mathbf{u}|\theta)$ is the probability that the vector of observed item scores for a person with trait level θ has the pattern $\mathbf{u}$, and $\mathrm{P}(u_i|\theta)$ is the probability that a person with trait level θ obtains a score of u_i on item i.

Similarly, the probability of the responses to a single item, i, by n individuals with abilities in the vector $\boldsymbol{\theta}$ is given by

$$P(\mathbf{U}_i = \mathbf{u}_i \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} P(u_{ij} \mid \theta_j) = P(u_{i1} \mid \theta_1) P(u_{i2} \mid \theta_2) \cdots P(u_{in} \mid \theta_n), \tag{2.3}$$

where $\mathbf{U_i}$ is the vector of responses to Item i for persons with abilities in the $\boldsymbol{\theta}$-vector, u_{ij} is the response on Item i by Person j, and θ_j is the trait level for Person j.

The property of local independence generalizes to the probability of the complete matrix of item responses. The probability of the full matrix of responses of n individuals to I items on a test is given by

$$P(\mathbf{U} = \mathbf{u} \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} \prod_{i=1}^{I} P(u_{ij} \mid \theta_j). \tag{2.4}$$

Although the assumptions of monotonicity and local independence are not necessary components of an item response theory, they do simplify the mathematics required to apply the IRT models. The monotonicity assumption places limits on the mathematical forms considered for the function[1] in (2.1), and the local independence assumption greatly simplifies the procedures used to estimate the parameters of the models.

The three assumptions that have been described above (i.e., one person parameter, monotonicity, and local independence) define a general class of IRT models. This class of models includes those that are commonly used to analyze the item responses from tests composed of dichotomously scored test items such as aptitude

[1] Nonmonotonic IRT models have been proposed (e.g., Thissen and Steinberg 1984, Sympson 1983), but these have not yet been generalized to the multidimensional case so they are not considered here.

and achievement tests. This class of models can be considered as a general psychometric theory that can be accepted or rejected using model checking procedures. The assumption of local independence can be tested for models with a single person parameter using the procedures suggested by Stout (1987) and Rosenbaum (1984). These procedures test whether the responses to items are independent when a surrogate for the person parameter, such as the number-correct score, is held constant. If local independence conditional on a single person parameter is not supported by observed data, then item response theories based on a single person parameter are rejected and more complex models for the data should be considered.

The general form of IRT model given in (2.1) does not include any specification of scales of measurement for the person and item parameters. Only one scale has defined characteristics. That scale is for the probability of the response to the test item that must range from 0 to 1. The specification of the function, f, must also include a specification for the scales of the person parameter, θ, and the item parameters, $\boldsymbol{\eta}$. The relative size and spacing of units along the θ-scale are determined by the selection of the form of the mathematical function used to describe the interaction of persons and items. That mathematical form sets the metric for the scale, but the zero point (origin) and the units of measurement may still not be defined. Linear transformations of a scale retain the same shape for the mathematical function.

For an IRT model to be considered useful, the mathematical form for the model must result in reasonable predictions of probabilities of all item scores for all persons and items in a sample of interest. The IRT model must accurately reflect these probabilities for all items and persons simultaneously. Any functional form for the IRT model will fit item response data perfectly for a one-item test because the locations of the persons on the θ-scale are determined by their responses to the one item. For example, placing all persons with a correct response above a point on the θ-scale and all of those with an incorrect response below that point and specifying a monotonically increasing mathematical function for the IRT model will insure that predicted probabilities are consistent with the responses. The challenge to developers of IRT models is to find functional forms for the interaction of persons and items that apply simultaneously to the set of responses by a number of persons to all of the items on a test.

The next section of this chapter summarizes the characteristics of several IRT models that have been shown to be useful for modeling real test data. The models were chosen for inclusion because they have been generalized to the multidimensional case. No attempt is made to present a full catalogue of UIRT models. The focus is on presenting information about UIRT models that will facilitate the understanding of their multidimensional generalizations.

2.1.1 Models for Items with Two Score Categories

UIRT models that are most frequently applied are those for test items that are scored either correct or incorrect – usually coded as 1 and 0, respectively. A correct response is assumed to indicate a higher level of proficiency than an incorrect response

so monotonically increasing mathematical functions are appropriate for modeling the interactions between persons and items. Several models are described in this section, beginning with the simplest. Models for items with two score categories (dichotomous models) are often labeled according to the number of parameters used to summarize the characteristics of the test items. That convention is used here.

2.1.1.1 One-Parameter Logistic Model

The simplest commonly used UIRT model has one parameter for describing the characteristics of the person and one parameter for describing the characteristics of the item. Generalizing the notation used in (2.1), this model can be represented by

$$P(U_{ij} = u_{ij} \mid \theta_j) = f(\theta_j, b_i, u_{ij}), \tag{2.5}$$

where u_{ij} is the score for Person j on Item i (0 or 1), θ_j is the parameter that describes the relevant characteristics of the jth person – usually considered to be an ability or achievement level related to performance on Item i, and b_i is the parameter describing the relative characteristics of Item i – usually considered to be a measure of item difficulty.[2]

Specifying the function in (2.5) is the equivalent of hypothesizing a unique, testable item response theory. For most dichotomously scored cognitive test items, a function is needed that relates the parameters to the probability of correct response in such a way that the monotonicity assumption is met. That is, as θ_j increases, the functional form of the model should specify that the probability of correct response increases. Rasch (1960) proposed the simplest model that he could think of that met the required assumptions. The model is presented below:

$$P(u_{ij} = 1 \mid A_j, B_i) = \frac{A_j B_i}{1 + A_j B_i}, \tag{2.6}$$

where A_j is the single person parameter now generally labeled θ_j, and B_i is the single item parameter now generally labeled b_i.

This model has the desired monotonicity property and the advantage of simplicity. For the function to yield values that are on the 0 to 1 probability metric, the product of $A_j B_i$ can not be negative because negative probabilities are not defined. To limit the result to the required range of probabilities, the parameters are defined on the range from 0 to ∞.

The scale for the parameters for the model in (2.6) makes some intuitive sense. A 0 person parameter indicates that the person has a 0 probability of correct response for any item. A 0 item parameter indicates that the item is so difficult that no matter

[2] The symbols used for the presentation of the models follow Lord (1980) with item parameters represented by Roman letters. Other authors have used the statistical convention of representing parameters using Greek letters.

how high the ability of the persons, they still have a 0 probability of correct response. In a sense, this model yields a proficiency scale that has a true 0 point and it allows statements like "Person j has twice the proficiency of Person k." That is, the scales for the model parameters have the characteristics of a ratio scale as defined by Stevens (1951).

Although it would seem that having a model with ratio scale properties would be a great advantage, there are also some disadvantages to using these scales. Suppose that the item parameter $B_i = 1$. Then a person with parameter $A_j = 1$ will have a .5 probability of correctly responding to the item. All persons with less then a .5 probability of correctly responding to the test item will have proficiency estimates that are squeezed into the range from 0 to 1 on the A-parameter scale. All persons with greater than a .5 probability of correct response will be stretched over the range from 1 to ∞ on the proficiency scale. If test items are selected for a test so that about half of the persons respond correctly, the expected proficiency distribution is very skewed. Figure 2.1 provides an example of such a distribution.

The model presented in (2.6) is seldom seen in current psychometric literature. Instead, a model based on a logarithmic transformation of the scales of the parameters (Fischer 1995a) is used. The equation for the transformed model is

$$P(u_{ij} = 1 \mid \theta_j, b_i) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} = \Psi(\theta_j - b_i), \tag{2.7}$$

where Ψ is the cumulative logistic density function, e is the base of the natural logarithms, and θ_j and b_i are the person and item parameters, respectively.

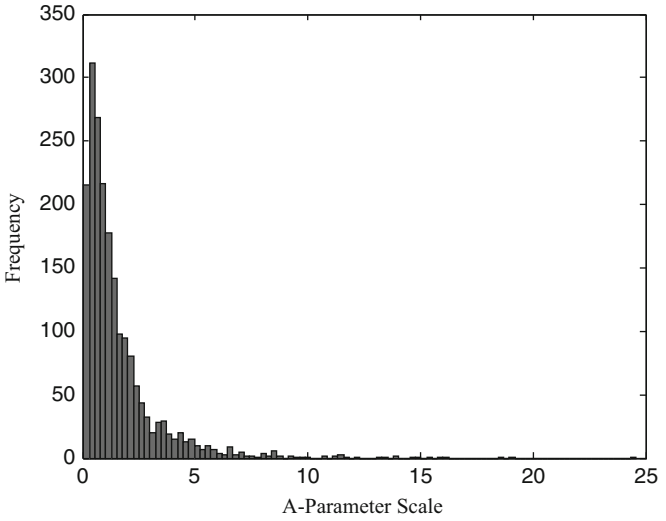

Fig. 2.1 Possible distribution of person parameters for the model in (2.6)

Because this model uses the logistic density function and because it has only a single item parameter, it is often called the one-parameter logistic IRT model. Alternatively, because it was originally suggested by Rasch (1960, 1961), it is called the *Rasch model*. The relationship between the models in (2.7) and (2.6) can easily be determined by substituting the transformations of the parameters into (2.7): $\theta_j = \ln(A_j)$ and $b_i = -\ln(B_i)$.

The scale of the person parameter in (2.7) ranges from $-\infty$ to ∞ rather than from 0 to ∞ for the model in (2.6). The scale for the item parameter is the same for that of the person parameter, but the direction of the scales are reversed from (2.6) to (2.7). Large values of the B_i parameter indicate easy items while small values of b_i indicate easy items. Thus, the b_i parameter is legitimately called a *difficulty parameter* while B_i is an "easiness" parameter.

The models in (2.6) and (2.7) show the flexibility that exists when hypothesizing IRT models. These two models will fit real test data equally well because the model parameters are a monotonic transformation of each other. Yet, the scales for the parameters are quite different in character. Equation (2.6) uses scales with a logical zero point while the parameters in (2.7) do not have that characteristic. It is difficult to say which model is a correct representation of the interactions between persons and an item. Generally, (2.7) seems to be preferred because estimation is more convenient and it has a clear relationship to more complex models.

Some researchers state that the Rasch model provides an interval scale of measurement using the typology defined by Stevens (1946). All IRT models have an interval scale for the person parameter as an unstated assumption because the form of the equation is not defined unless the scale of the person parameter has interval properties. But, in (2.6) and (2.7) we have two interval scales that are nonlinear transformations of each other. This would contradict the permissible transformations allowed by Stevens (1946). The point is that these scales are arbitrary decisions of the model builder. The usefulness of the scales comes from the ease of interpretation and the relationships with other variables, not inherent theoretical properties.

The form of (2.7) is shown graphically in Fig. 2.2. The graph shows the relationship between θ and the probability a person at that trait level will provide a correct response for an item with b-parameter equal to .5. The graph clearly shows the monotonically increasing relationship between trait level and probability of correct response from the model. It also shows that this model has a lower asymptote for the probability of 0 and an upper asymptote of 1. The graph of the probability of a correct response as a function of θ is typically called an *item characteristic curve* or ICC.

Several characteristics of this model can be derived through some relatively simple analysis. A cursory look at the graph in Fig. 2.2 shows that the curve is steeper (i.e., has greater slope) for some values of θ then for others. When the slope is fairly steep, the probability of correct response to the item is quite different for individuals with θ-values that are relatively close to each other. In regions where the ICC is fairly flat – has low slope – the θ-values must be relatively far apart before the probability of correct response results in noticeable change.

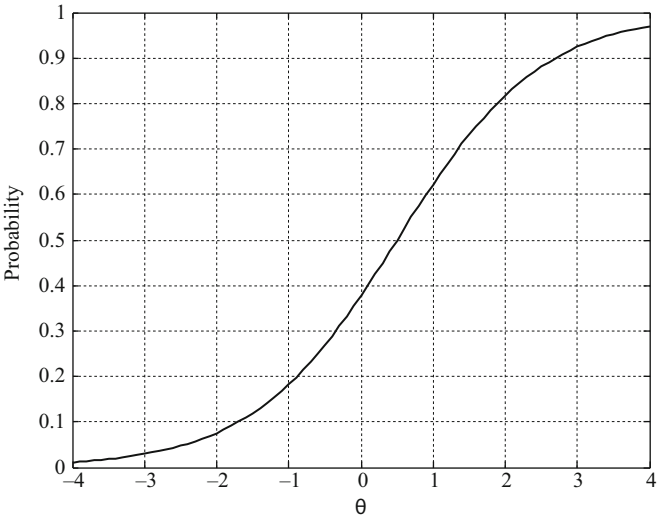

Fig. 2.2 One-parameter logistic model item characteristic curve for $b = .5$

For example, a person with $\theta = .25$ has a probability of correct response to the test item of .44 if the model is correct. A person .5 units higher on the θ-scale (i.e., $\theta = .75$) has a probability of correct response of .56 if the model is correct. The difference in these two probabilities for the .5-unit change on θ is .12. At $\theta = -1.5$, the probability of correct response is .12. At $\theta = -1.0$, a .5 unit change from the original value, the probability of correct response is .18. Over this .5 unit change in the θ-scale, the change in probability is only .06. This analysis indicates that the test item described by the ICC in Fig. 2.2 would be better at differentiating between persons between .25 and .75 on the θ-scale than persons between θ-values of -1.5 and -1.0.

A more analytic way of considering this issue is to determine the slope of the ICC at each point on the θ-scale so that the steepness of the ICC can be determined for any value of θ. The first derivative of the function describing the interaction of the persons and the item provides the slope of the ICC at each value of θ. For the model given in (2.7), the first derivative of $P(u_{ij} = 1 \mid \theta_j, b_i)$ with respect to θ is given by the following expression:

$$\frac{\partial P}{\partial \theta} = P - P^2 = P(1 - P) = PQ, \tag{2.8}$$

where, for simplicity,

$$P = P(u_{ij} = 1 \mid \theta_j, b_i),$$

and $Q = (1 - P)$.

From this expression, it is clear that the slope of the ICC is 0 (i.e., the curve is horizontal) only when the probability of correct response is 0 or 1. This confirms

that the asymptotes for the model are 0 and 1. Note that the slope of the ICC is equal to .25 (that is, $1/4$, a point that will be important later) when the probability of correct response is .5. A probability of .5 is the result of the model when θ_j is equal to b_i in (2.7) because then the exponent is equal to 0 and $e^0 = 1$ yielding a probability of $1/2$.

Note that the difference in probability of correct response for θ values of .25 and .75 (centered around $\theta = .5$) was .12. The slope at $\theta = .5$ can be approximated as the ratio of the difference in probability (.12) over the range of θs to the difference in θ values $(.5) - .12/.5 = .24$. Because this ratio gives the slope for a linear function and the ICC is nearly linear in the θ-range from .25 to .75, it is not surprising that the slope from (2.8) and the approximation have nearly the same value. The similarity merely confirms that the derivative with respect to θ provides the slope.

To determine where along the θ-scale, the test item is best at differentiating people who are close together on the scale, the derivative of the expression for the slope, (2.8), can be used. This is the second derivative of (2.7) with respect to θ. Setting the second derivative to 0 and solving for θ yields the point on the θ-scale where the test item is most discriminating – the point of maximum slope. The derivative of (2.8) with respect to θ is

$$\frac{\partial(P - P^2)}{\partial\theta} = (P - P^2)(1 - 2P). \tag{2.9}$$

Solving this expression for 0 yields values of P of 0, 1, and .5. The values of θ that correspond to the values of 0 and 1 are $-\infty$ and ∞, respectively. The only finite solution for a value of θ corresponds to the value of $P = .5$. That is, the slope of the function is steepest when θ has a value that results in a probability of correct response from the model of .5. As previously noted, a .5 probability results when θ is equal to b_i. Thus, the b-parameter indicates the point on the θ-scale where the item is most discriminating. Items with high b-parameters are most discriminating for individuals with high trait levels – those with high θ-values. Items with low b-parameters are most discriminating for individuals with low trait levels.

The b-parameter also provides information about the ranges of trait levels for persons that are likely to respond correctly or incorrectly to the test item. Those persons with θ-values greater than b_i have a greater than .5 probability of responding correctly to Item i. Those with θ-values below b_i have less than a .5 probability of responding correctly to the item. This interpretation of the b-parameter is dependent on the assumption that the model in (2.7) is an accurate representation of the interaction of the persons and the test item.

Because the b-parameter for the model provides information about the trait level that is best measured by the item and about the likely probability of correct response, it has been labeled the *difficulty parameter* for the item. Although this terminology is somewhat of an oversimplification of the complex information provided by this parameter, the label has been widely embraced by the psychometric community.

It is important to note that for the model given in (2.7), the value of the maximum slope is the same for all items. The only feature of the ICC that changes from test item to test item is the location of the curve on the θ-scale. Figure 2.3 shows ICCs

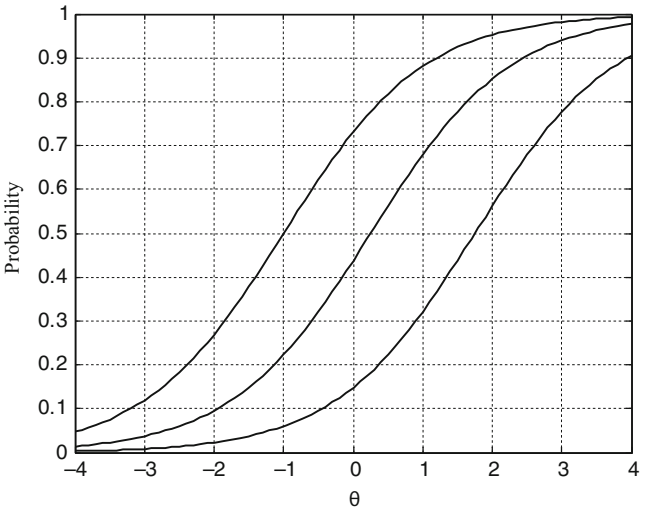

Fig. 2.3 ICCs for the model in (2.7) for $b_i = -1, .25, 1.75$

for three different items. The items have b-parameters with values of 1.75, .25, and -1.0. The graphs in Fig. 2.3 show that the slopes of the three curves are the same where they cross the $P = .5$ line. Because the ICCs for the model have a common value for the maximum slope for the items, the model is said to include an assumption that all test items have the same discriminating power. This is not true in an absolute sense because at different values of θ the slopes of the various items differ. But it is true that a set of items that are consistent with this model will have the same maximum slope, and the magnitude of the maximum slope at $P = .5$ will be $1/4$ for all items.

The one-parameter logistic (Rasch) model has a number of advantages over more complex models including simplicity in form and mathematical properties that make the estimation of the parameters of the model particularly convenient. One convenient mathematical property is that there is a direct relationship between the number-correct scores for a set of test items and the estimates of θ. All persons with the same number-correct score have the same maximum-likelihood estimate of θ.[3] A similar relationship exists between the number of correct responses to a test item and the maximum-likelihood estimate of the b-parameter for the test item. All items with the same proportion of correct responses for a sample of examinees have the same maximum-likelihood b-parameter estimate based on that sample. These relationships allow the θ-parameters for the examinees to be estimated independently of the b-parameters for the test items. A significant contingent of the psychometric

[3] Chapter 6 presents a number of estimation procedures including maximum likelihood. A full discussion of estimation procedures is beyond the scope of this book. The reader should refer to a comprehensive mathematical statistics text for a detailed discussion of maximum-likelihood estimation and other techniques for estimating model parameters.

community maintains that these properties of the one-parameter (Rasch) model are so desirable that models that do not have these properties should not be considered. The perspective of this group is that only when person and item-parameters can be estimated independently of each other do θ-estimates result in numbers that can be called measurements. Andrich (2004) provides a very clear discussion of this perspective and contrasts it with alternative views.

The perspective taken here is that the goal of the use of IRT models is to describe the interaction between each examinee and test item as accurately as possible within the limitations of the data and computer resources available for test analysis. This perspective is counter to the one that proposes that strict mathematical criteria are needed to define measurement and only models that meet the criteria are acceptable. Rather, the value of a model is based on the accuracy of the representation of the interaction between persons and items. The estimates of parameters of such models are used to describe the characteristics of the persons and items. The strict requirements for independent estimation of person and item parameters will not be considered as a requirement for a useful psychometric model.

From this perspective, whether or not the one-parameter logistic model is adequate for describing the interaction of examinees with test items is an empirical question rather than a theoretical question. For example, when test forms are analyzed using traditional item analysis procedures, the usual result is that the point-biserial and biserial correlation indices of the discriminating power of test items vary across items. Lord (1980) provides an example of the variation in discrimination indices for data from a standardized test. Item analysis results of this kind provide strong evidence that test items are not equal in discriminating power. These results imply that a model that has the same value for the maximum slope of the ICCs for all test items does not realistically describe the interaction between persons and items.

2.1.1.2 Two-Parameter Logistic Model

Birnbaum (1968) proposed a slightly more complex model than the one presented in (2.7). The mathematical expression for the model is given by

$$P(U_{ij} = 1 \mid \theta_j, a_i, b_i) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{2.10}$$

where a_i is a parameter related to the maximum slope of the ICC and the other symbols have the same definition as were given for (2.7). The first partial derivative of this model with respect to θ_j, which gives the slope of the ICC at each value of θ_j, is given by

$$\frac{\partial P(U_{ij} = 1 \mid \theta_j, a_i, b_i)}{\partial \theta_j} = a_i P_2 Q_2 = a_i (P_2 - P_2^2), \tag{2.11}$$

where P_2 is the probability of a correct response for the model given in (2.10), and $Q_2 = (1 - P_2)$. The subscript "2" is used to distinguish the probability estimates from the model in (2.10) from those for the model in (2.7).

The partial derivative of the slope with respect to θ_j, which is the same as the second partial derivative of the model in (2.10) with respect to θ_j, is given by

$$\frac{\partial(a_i P_2 Q_2)}{\partial \theta_j} = a_i^2(P_2 - P_2^2)(1 - 2P_2). \tag{2.12}$$

Solving this expression for zero to determine the point of maximum slope shows that the only finite solution occurs when $P_2 = .5$, as was the case for the one-parameter logistic model. Substituting .5 into (2.11) gives the value of the maximum slope, $a_i/4$. This result shows that the a_i parameter controls the maximum slope of the ICC for this model. It should also be noted that the point of maximum slope occurs when $\theta_j = b_i$, just as was the case for the one-parameter logistic model. Thus, b_i in this model can be interpreted in the same way as for the one-parameter logistic model – as the difficulty parameter.

Figure 2.4 presents three examples of ICCs with varying a- and b-parameters. These ICCs are not parallel at the point where they cross the .5 probability line. Unlike the ICCs for the one-parameter logistic model, these curves cross each other. This crossing of the curves allows one test item to be harder than another test item for persons at one point on the θ-scale, but easier than that same item for persons at another point on the θ-scale. For example at $\theta = -1$, Item 1 is easier than Item 2 (the probability of correct response is greater), but at $\theta = 1$, Item 1 is more

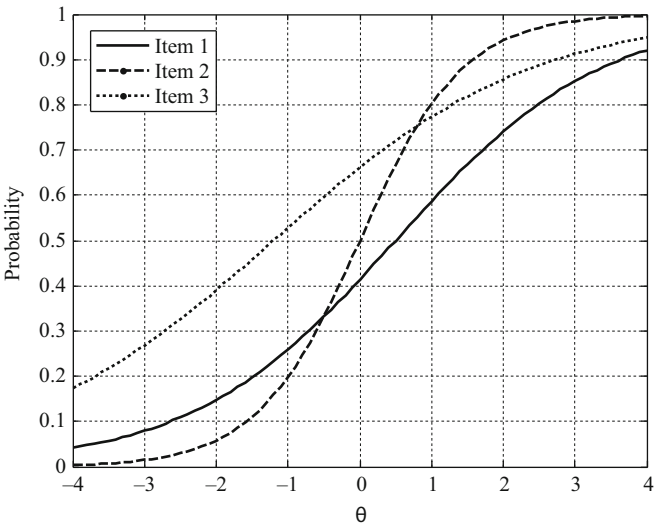

Fig. 2.4 Item characteristic curves for the two-parameter logistic model with a-parameters .7, 1.4, .56 and b-parameters .5, 0, −1.2, respectively