Chapter 4
Multidimensional Item Response Theory Models

As the previous chapters suggest, it is not difficult to conceive of test items that require more than one hypothetical construct to determine the correct response. However, when describing multidimensional item response theory (MIRT) models, care should be taken to distinguish between dimensions as defined by MIRT models, which represent statistical abstractions of the observed data, and the hypothetical constructs that represent cognitive or affective dimensions of variation in a population of examinees. The earlier chapters present some of those distinctions. This chapter will elaborate on the distinctions between coordinates and constructs and the distinctions will be given additional treatment in Chaps. 6 and 7.

There are two major types of multidimensional item response theory models. The types are defined by the way the information from a vector of θ-coordinates is combined with item characteristics to specify the probability of responses to the item. One type of model is based on a linear combination of θ-coordinates. That linear combination is used with a normal ogive or logistic form to specify the probability of a response. The linear combination of θ-coordinates can yield the same sum with various combinations of θ-values. If one θ-coordinate is low, the sum will be the same if another θ-coordinate is sufficiently high. This feature of this type of model has been labeled as compensation and models with this property are typically called *compensatory models* to emphasize that property. Because of its common use, that terminology will be used here as a short-hand description for that type of model.

The second type of model separates the cognitive tasks in a test item into parts and uses a unidimensional model for each part. The probability of correct response for the item is the product of the probabilities for each part. The use of the product of probabilities results in nonlinear features for this class of models. Also, the fact that the probability of correct response can not exceed the highest of the probabilities in the product reduces the compensation of a high θ-coordinate for a low θ-coordinate. These models are often called *noncompensatory models* in the MIRT literature, but the term *partially compensatory* will be used here because a high θ-coordinate on one dimension does yield a higher probability of response than a low value on that dimension. Therefore, some compensation does occur.

Within these two major types of MIRT models, there are a number of model variations. This chapter describes both of these model types and their characteristics. The type based on the linear combination of θ-coordinates (compensatory models)

M.D. Reckase, *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-89976-3_4,

is described first because of its close connection to factor analysis and because models of this type are more prevalent in the research literature.

A major component of variation within MIRT model type is the number of possible score points for the test items that are being modeled. The historic antecedents of MIRT presented in Chap. 3 dealt solely with models for items with two response categories. More recent work (e.g., Adams et al. 1997; Kelderman and Rijkes 1994; Muraki and Carlson 1993; Yao and Schwarz 2006) has extended unidimensional models for test items with more than two score categories to the multidimensional case. At this time, the MIRT models for polytomous items all fall within the category of compensatory models, but it is certainly possible that partially compensatory models for more than two score categories will be developed.

To provide a context for the MIRT models that will be presented in this chapter, two examples of dichotomously scored test items are provided. For these examples, it is assumed that all of the test items written for a test depend strictly on two underlying skill constructs that are labeled (1) arithmetic problem solving, and (2) algebraic symbol manipulation. These test items might also be considered to require skill in reading English, but that construct will be ignored for these examples to allow simple tabular presentation of the examples.

The fact that the items require capabilities on two constructs to determine the correct response to the items suggests that a trait space with two coordinate axes is needed to describe the variation in examinee responses to the set of items on the test. The coordinates for a specific examinee j for this space are indicated by $(\theta_{j1}, \theta_{j2})$. Further, the function P $(\theta_{j1}, \theta_{j2})$ will be used to indicate the probability of correct response to an item given the location of examine j in the space. In general, the coordinates $(\theta_{j1}, \theta_{j2})$ need not be associated with any hypothetical constructs, such as those listed earlier. Under some circumstances, however, it may be possible to align the coordinate axes with the hypothetical constructs. In such cases, estimation of the location of an examinee will also provide estimates of the levels on the constructs of interest. For example, θ_{j1} could be an estimate of the examinee's level on arithmetic problem solving and θ_{j2} could be an estimate of the examinee's level on algebraic symbol manipulation. For most of the MIRT models, estimates of the constructs range from $-\infty$ to $+\infty$ along the coordinate axes. Larger values indicate a greater capability on the construct than smaller values.

An example test item that requires both of the hypothesized constructs is given below.

1. A survey asked a sample of people which of two products they preferred. 50% of the people said they preferred Product A best, 30% said they preferred Product B, and 20% were undecided. If 1,000 people preferred Product A, how many people were undecided?

   A. 200
   B. 400
   C. 800
   D. 1,200
   E. 2,000

Table 4.1 Proportions of correct responses to Item 1 for 4,114 examinees

Midpoints θ_{j1}	Midpoints of θ_{j2} intervals							
	−1.75	−1.25	−.75	−.25	.25	.75	1.25	1.75
−1.75								
−1.25	.20		.09					
−.75	.06	.18	.39	.47	.19	.67		
−.25	.18	.25	.30	.45	.54	.50	.61	.82
.25	.19	.40	.39	.53	.45	.46	.77	.57
.75	.24	.34	.49	.53	.50	.65	.76	.71
1.25	.30	.35	.54	.55	.47	.63	.78	.55
1.75	.51	.55	.57	.62	.60	.71	.71	.65

The empty cells have frequencies of less than 10 so proportions were not computed for those cells

An examination of this test item might suggest that it requires some level of proficiency on both hypothetical constructs. It clearly requires some arithmetic computation. It could also require some algebraic symbol manipulation because the examinees must solve for some unknown values. If some level of skill on both constructs is required to determine the correct solution, then differences in responses of examinees with various combinations of knowledge or skill for the two constructs would be expected.

Suppose this test item is administered to a large number of examinees with known θ coordinates. The empirical proportions of correct response for the test item at $(\theta_{j1}, \theta_{j2})$ points can be calculated. Table 4.1 presents the proportions of 4,114 examinees who responded correctly to a single test item. These data are from a random sample of examinees who have taken a college entrance examination in the United States. Eight intervals are used to summarize the θ-coordinates. Each interval is .5-unit wide. The midpoints of the intervals are shown in the table. The proportion correct values for combinations of θ-coordinates that had frequencies of less than 10 are not shown. Overall, the tabled results present a pattern that as the values of either θ-coordinate increases, the proportion of correct responses tend to increase. Furthermore, the proportion of correct responses tends to be low when both θ_1 and θ_2 are low and high when both are high. There are also combinations of θ_1 and θ_2 that give approximately the same proportions of correct response. For example, if θ_1 is 1.75 and θ_2 is −1.75, the proportion of correct responses is .51. Similarly, if θ_1 is −.25 and θ_2 is .75, the proportion of correct responses is .50. A series of cells from lower left to upper right tend to have the same proportion of correct responses.

The classical test theory statistics for this item are a proportion correct of .54 and a point-biserial correlation with the total score of .49. A unidimensional IRT analysis with the three-parameter logistic model yields an item discrimination parameter estimate of 1.11, difficulty parameter estimate of .01, and pseudo-guessing parameter estimate of .08. Also, the unidimensional model did not yield good fit to the item response data from this test item. The chi-square fit statistic from the BILOG MG program (Zimowski et al. 2003) was 60.4 with 8 degrees of freedom. This is not a

surprising result because the test item violates the assumption that only a single trait is needed to determine the correct response.

The general pattern of the response function for Item 1, $\mathbf{P}_1(\theta_1, \theta_2)$, tends to follow a monotonically increasing pattern of proportions of correct response across the $\boldsymbol{\theta}$-space. Further, the fairly consistent pattern of increasing proportions conditional on θ_1 and θ_2 shows that the test item is sensitive to differences on each coordinate dimension. If those coordinate dimensions align with the hypothetical constructs arithmetic problem solving and algebraic symbol manipulation, the test item is sensitive to differences on each of the constructs.

Consider a second test item that involves only a small amount of the hypothetical construct arithmetic problem solving, but requires a substantial amount of the hypothetical construct algebraic symbol manipulation. An example test item of this type is:

2. For all x, $(2x + 3)^2 + 2(2x + 4) - 2$ equals which of the following expressions?
   A. $4x^2 + 4x + 11$
   B. $(4x + 15)(x + 1)$
   C. $(2x + 5)(2x + 3)$
   D. $(2x + 5)(2x + 2)$
   E. $(2x + 5)(2x - 1)$.

Table 4.2 presents the proportions of correct response for Item 2 for the same 4,114 examines that responded to Item 1. For this test item, the proportion of correct responses increases quite dramatically with an increase in θ_2, but there is little change with an increase in θ_1. The test item is relatively insensitive to changes on one coordinate dimension, while it is very sensitive to changes on the other coordinate dimension. The overall proportion correct for this test item is .38 and it has a point-biserial correlation with the number-correct score of .26. The unidimensional IRT parameter estimates for this test item are $a = 1.175$, $b = .933$, and $c = .12$. As with the previous test item, the unidimensional IRT model did not fit the item response data for this item very well. The chi-square goodness of fit statistic had a value of 68.4 with 8 degrees of freedom. This statistic indicates that the probability

Table 4.2 Proportion of correct responses to Item 2 for 4,114 examinees

Midpoints θ_{j1}	Midpoints of θ_{j2} intervals							
	−1.75	−1.25	−.75	−.25	.25	.75	1.25	1.75
−1.75								
−1.25	.10		.18					
−.75	.06	.00	.22	.33	.31	.40		
−.25	.14	.18	.36	.28	.25	.36	.28	.91
.25	.10	.19	.18	.21	.29	.37	.65	.86
.75	.20	.29	.30	.37	.31	.43	.55	.88
1.25	.11	.11	.23	.36	.43	.43	.59	.97
1.75	.07	.15	.15	.19	.36	.60	.76	.99

The empty cells have frequencies of less than 10, so proportions were not computed for those cells

was well below .001 that the data were generated from the unidimensional model. The poor fit is present even though the item mainly assesses the single construct of ability to perform algebraic manipulations. The lack of fit is due to the fact that the construct measured by the test item does not match the dominant construct measured by the full set of test items.

The unidimensional discrimination parameter estimates for these two test items are approximately the same, but they differ in difficulty. However, more than a shift in difficulty is indicated by the results in the two tables. A comparison of two cells in Tables 4.1 and 4.2 illustrates the differences. In Table 4.1, cells (1.75, −1.75) and (−.25, .75) had proportions correct of approximately .5. In Table 4.2, the corresponding values are .07 and .36. These results show that the interactions of the two items with the locations of persons in the coordinate space are quite different. In Sect. 4.1 of this chapter, these differences will be shown to indicate that the test items are sensitive to different combinations of skill dimensions along with having different difficulty.

The estimated proportions of correct response in Tables 4.1 and 4.2 provide two types of information. First, they give approximations of the probability of correct response for examinees with locations given by specified $\boldsymbol{\theta}$-vector. The values in the table could be smoothed and probabilities could be estimated by interpolation for any $\boldsymbol{\theta}$-vector. Second, the pattern of increase in the proportions for each test item indicates the general characteristics of the item for assessing the constructs that define the coordinate axes. Some of those characteristics are the direction in the $\boldsymbol{\theta}$-space that yields the greatest increase in proportion of correct response for a change of location in that direction and information about the difficulty of the test item for examinees at locations specified by a $\boldsymbol{\theta}$-vector. The direction of greatest increase in the space indicates the sensitivity of the test item to changes in levels of the hypothesized constructs. These characteristics suggest possible parameters for a function that models the relationship between the coordinates for an examinee and the probability of correct response to the test item.

In MIRT, the intent is to specify a model that provides a reasonable representation of data like that presented in Tables 4.1 and 4.2 and estimate the parameters of the model. The term *structural* is used to describe the parameters that describe the functioning of the test items. The term *incidental* is used to describe the vector of coordinates describing the locations of individuals (Hambleton and Swamination 1983; Neyman and Scott 1948). In this text, the vector, $\boldsymbol{\theta}$, represents the incidental parameters, and Roman letters are used to represent the structural parameters.

The mathematical function chosen as the MIRT model is fit to the observed proportions of correct response. When the fit is reasonably good, the resulting model provides estimates of the conditional probability of correct response given the coordinates $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ for the m-dimensional space, where m is the number of dimensions used to model the data. The MIRT model is assumed to be a continuous probability function relating the location specified by $\boldsymbol{\theta}$ to the probability of correct response to a test item i with specified structural parameters. This model can be represented in the following ways for a test item scored 1 for a correct response and 0 for an incorrect response.

$$P_i(\theta_1, \ldots, \theta_m) = \Pr(u_i = 1|\theta_1, \ldots, \theta_m) \equiv P_i(\boldsymbol{\theta}). \tag{4.1}$$

In these expressions, the subscript i indicates the item and there is an implied set of structural parameters for the item, even though they are not explicitly listed. In the next section of this chapter, the structural parameters will be defined for several different models.

A more general representation of a MIRT model is given in (4.2). In this equation, $\boldsymbol{\eta}$ represents a vector of structural parameters that describe the characteristics of the test item, U represents the score on the test item and u is a possible value for the score, and f is a function that describes the relationship between the locations of persons specified by $\boldsymbol{\theta}$ and the probability of the response.

$$P_i(U = u|\boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{\eta}_\mathbf{i}, u). \tag{4.2}$$

The item score, u, may appear on both sides of the equation if the test item is scored either correct (1) or incorrect (0) to change the form of the function depending on the value of the score. When more than two score categories are present for the item, this mathematical convenience is not used.

Most MIRT models assume that the probability of selecting or producing the correct response to a test item scored as either correct or incorrect increases as any element in the $\boldsymbol{\theta}$-vector 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 affect the response to an item produced by another person. Also, the response by a person to one item does not change 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 upon the person's $\boldsymbol{\theta}$-vector and the item's vector of parameters, $\boldsymbol{\eta}$. The practical implications of these assumptions are that the testing process needs to be controlled so that examinees do not share information during the test, and that tests must be constructed so that information in one test item does not increase or decrease 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 $\boldsymbol{\theta}$-vector, but the assumption does not generalize to the case of variation in $\boldsymbol{\theta}$-elements. For groups of individuals with variation in the constructs 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 MIRT model hold, the correlation between item scores will be due to variation in elements in the person parameter vector.

Because of the local independence assumption, the probability of a collection of responses (responses from one person to the items on a test or the responses from 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 vector $\boldsymbol{\theta}$ is the product of the probabilities of the individual responses, u_i, to the items on an I-item test.

$$P(U_1 = u_1, \ldots U_I = u_I|\boldsymbol{\theta}) = \prod_{i=1}^{I} P(U_i = u_i|\boldsymbol{\theta}). \tag{4.3}$$

McDonald (1967, 1981) indicates that sometimes a weaker assumption than (4.3) can be used to develop estimation procedures for the models. The weaker assumption is that the conditional covariances between all pairs of items are zero. That is

$$E(\text{cov}(U_i, U_j)|\boldsymbol{\theta}) = 0, \quad i \neq j, \tag{4.4}$$

for all values of the $\boldsymbol{\theta}$-vector.

The following sections of this chapter will describe the characteristics of individual MIRT models. These models use a number of functional forms and they include models for both dichotomously and polytomously scored items.

4.1 Multidimensional Models for the Interaction Between a Person and a Test Item

As was the case for UIRT models, MIRT comprises a set of models (item response theories), which have as a basic premise that the interaction between a person and a test item can be modeled reasonably accurately by a specific mathematical expression. Many different mathematical expressions have been developed for MIRT models. Some of them have already been presented in Chap. 3. This section will provide descriptions of the MIRT models that most commonly appear in the research literature. The MIRT models for items with two score categories will be presented first. These models have a relatively long history in the psychometric literature, and there is more experience with their application. The models for items with more than two score categories are described next. Chapter 5 presents statistical ways for representing characteristics of test items that are sensitive to differences on multiple dimensions. These statistical measures parallel those presented for UIRT models in Chap. 2.

4.1.1 MIRT Models for Test Items with Two Score Categories

Some of the stimulus for the development of MIRT came from attempts to addressing the problem of factor analyzing dichotomous data. For that reason, MIRT models for dichotomous items have appeared in the research literature since the 1980s (e.g., Bock and Aitken 1981). Because of the importance of this early work, the MIRT models for dichotomous items (those with two score categories) are

presented first, beginning with the model that is an extension of the two-parameter logistic UIRT model. That model is used to show basic MIRT concepts. Then, alternatives to that model with the same basic form are presented. These include models based on the normal ogive rather than the logistic function and models that have fewer or greater numbers of item parameters. A different extension of UIRT models is presented next, followed by comparisons of the different types of models.

4.1.1.1 Compensatory Extensions of the UIRT Models

Multidimensional extension of the two-parameter logistic model. The two-parameter logistic model (see (2.10)) has an exponent of the form $a(\theta - b)$. Multiplying through by a results in $a\theta - ab$. If $-ab$ is replaced by d, the expression is in what is called slope/intercept form, $a\theta + d$. One way of extending the two-parameter logistic model to the case where there are multiple elements in the $\boldsymbol{\theta}$-vector is to replace the simple slope/intercept form with the expression $\mathbf{a}\boldsymbol{\theta}' + d$, where $\mathbf{a}$ is a $1 \times m$ vector of item discrimination parameters and $\boldsymbol{\theta}$ is a $1 \times m$ vector of person coordinates with m indicating the number of dimensions in the coordinate space. The intercept term, d, is a scalar. The form of the multidimensional extension of the two-parameter logistic (M2PL) model is given by

$$P(U_{ij} = 1|\boldsymbol{\theta}_\mathbf{j}, \mathbf{a}_\mathbf{i}, d_i) = \frac{e^{\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i}}{1 + e^{\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i}}. \tag{4.5}$$

The exponent of e in this model can be expanded to show the way that the elements of the $\mathbf{a}$ and $\boldsymbol{\theta}$ vectors interact.

$$\mathbf{a_i}\boldsymbol{\theta_j}' + d_i = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \cdots + a_{im}\theta_{jm} + d_i = \sum_{\ell=1}^{m} a_{i\ell}\theta_{j\ell} + d_i. \tag{4.6}$$

The exponent is a linear function of the elements of $\boldsymbol{\theta}$ with the d parameter as the intercept term and the elements of the $\mathbf{a}$-vector as slope parameters. The expression in the exponent defines a line in an m-dimensional space. This results in an interesting property for this model. If the exponent is set to some constant value, k, all $\boldsymbol{\theta}$-vectors that satisfy the expression $k = \mathbf{a_i}\boldsymbol{\theta_j}' + d_i$ fall along a straight line and they all yield the same probability of correct response for the model.

This relationship can be shown graphically if the number of coordinate axes is assumed to be two. For example, suppose that $k = 0$. For this value of the exponent, the probability of correct response is .5 because $e^0 = 1$. The expression on the right of (4.5) simplifies to $1/2$. For a test item with $\mathbf{a}$-vector equal to [.75 1.5] and d-parameter equal to $-.7$, the exponent of the model for this item is $.75\theta_1 + 1.5\theta_2 - .7 = 0$. Rearranging terms to put this expression into the usual slope/intercept form for a line results in

$$\theta_2 = -.5\theta_1 + \frac{.7}{1.5}. \tag{4.7}$$

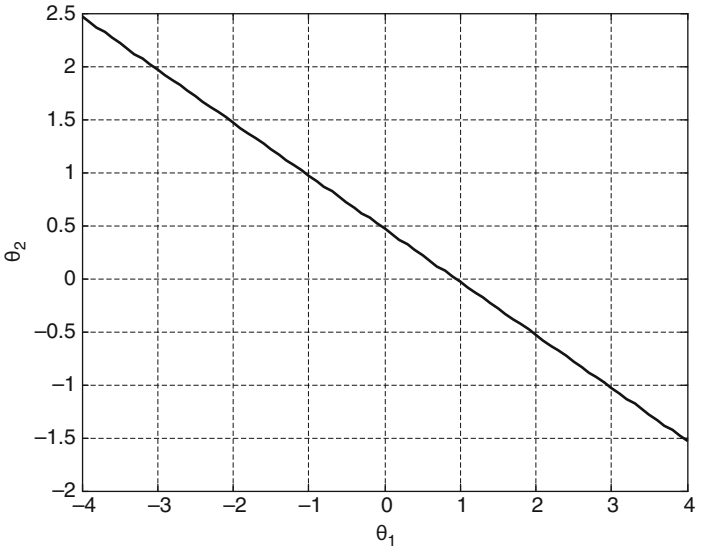

Fig. 4.1 Plot of θ-vectors that yield exponents of $k = 0$ for a test item with parameters $a_1 = .75$, $a_2 = 1.5$, $d = -.7$

This is the equation for a line with slope $-.5$ and intercept $.7/1.5 = .467$. This line is shown in Fig. 4.1.

This plot shows an interesting property of this MIRT model. The model indicates that all persons with $\boldsymbol{\theta}$-vectors that fall on the line have a probability of correct response of .5. Persons with $\boldsymbol{\theta}$-vectors of [0 .5], moderate values on both sets of coordinates, and [−4 2.5], a very low value on the first coordinate and a high value on the second coordinate, have predicted probabilities of .5. When the coordinates are interpreted as abilities, this feature of the model has been used to indicate that a high ability on one dimension can compensate for a low ability on another dimension. This is also shown by the fact that a $\boldsymbol{\theta}$-vector of [4 −1.5] also falls on the line yielding a probability of correct response of .5. Because of this feature of the model, it has been labeled as a *compensatory* MIRT model. Of course, the compensatory nature of the model also holds for higher dimensional cases. Low values of any coordinate, or any combinations of coordinates, can be compensated for by a high value on another coordinate, or combinations of coordinates, to yield the same probability of response as more moderate values.

Graphs of the model for the two-dimensional case clearly show the linear form of the exponent and the compensatory nature of the model. Figure 4.2 shows the form of the model in two ways. The left panel shows the probability of correct response to the item as the height above the (θ_1, θ_2)-plane. This shows the item response surface (IRS) for the item. The right panel shows the probabilities as contours of the surface shown in the left panel. The example uses the same item parameters as used in Fig. 4.1.

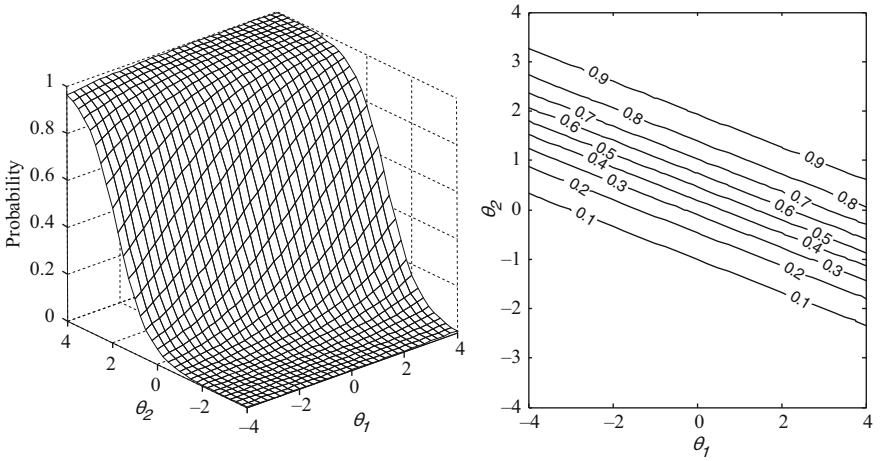

Fig. 4.2 Surface plot and contour plot for probability of correct response for an item with $a_1 = .5$, $a_2 = 1.5$, $d = -.7$

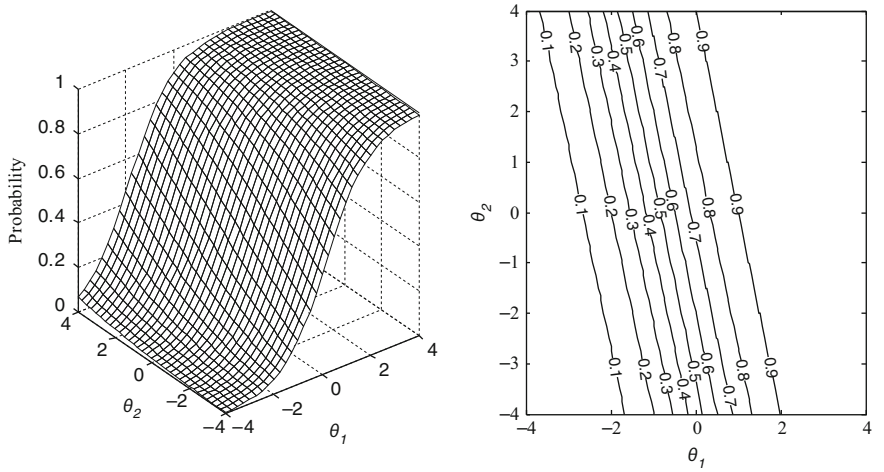

Fig. 4.3 Surface plot and contour plot for the probability of correct response for an item with $a_1 = 1.2$, $a_2 = .3$, $d = 1$

The plots show the way the model represents the characteristics of the test item. Both representations show that the probability of correct response increases monotonically with an increase in one or both elements of the $\boldsymbol{\theta}$-vector. Further, it is clear that the contours of equal probability form straight lines. These lines show the compensatory nature of the model. The plots also show that the probability of correct response increases more quickly with changes in location parallel to the θ_2-axis than changes parallel to the θ_1-axis. The rate of change of the probabilities with changes in location corresponds to the difference in the respective a-parameters. For comparison purposes, Fig. 4.3 shows the corresponding plots for an item with $a_1 = 1.2$,

$a_2 = .3$, and $d = 1.0$. The orientation of the surface for the second test item is quite different than for the first test item because the second test item increases in probability more quickly along the θ_1 coordinate axis and the first test item increases more quickly along the θ_2 coordinate axis.

Some intuitive meaning for the parameters of the model can be determined from careful inspection of Figs. 4.2 and 4.3. First, the scales for θs are from -4 to 4 for the plots. In the standard forms of the models, θs generally fall in this range, but the actual range is from $-\infty$ to ∞ for each coordinate dimension. The zero point and unit of measurement for each coordinate dimension is arbitrary. The values of these features of the θs are usually constrained to be the mean and standard deviation of the sample first used to calibrate a set of test items, but they can be transformed to match other characteristics of the sample of examinees or the characteristics of the items. The possible transformations of parameters will be described in Chap. 8.

The a-parameters indicate the orientation of the equiprobable contours and the rate that the probability of correct response changes from point to point in the $\boldsymbol{\theta}$-space. This can be seen by taking the first partial derivative of the expression in (4.5) with respect to a particular coordinate dimension, θ_ℓ. To simplify the presentation of the results $\mathrm{P}(U_{ij} = 1|\boldsymbol{\theta}_\mathbf{j}, \mathbf{a_i}, d_i) = \mathrm{P}$ and $\mathrm{Q} = (1 - \mathrm{P})$.

$$\frac{\partial P}{\partial \theta_\ell} = a_\ell P(1 - P) = a_\ell PQ. \tag{4.8}$$

This result shows that the slope of the IRS parallel to a coordinate axis has the same form as for the two-parameter logistic model shown in (2.11). The slope is greatest, $1/4\, a_\ell$, when the probability of correct response is .5. Because the a-parameters are related to the slope of the surface and the rate of change of the probability with respect to the coordinate axes, the $\mathbf{a}$-parameter is usually called the *slope* or *discrimination parameter*.

The probability of correct response for a test item is .5 when the exponent in (4.5) is 0. That is $\mathrm{e}^0/(1 + \mathrm{e}^0) = 1/(1 + 1) = 1/2$. When the exponent of e is 0, the exponent takes the form $\mathbf{a_i}\boldsymbol{\theta}'_\mathbf{j} + d_i = 0$. This equation is the expression for the line in the $\boldsymbol{\theta}$-space that describes the set of locations in the space that have a .5 probability of a correct response. If all of the elements of $\boldsymbol{\theta}$ are equal to 0 except one, say θ_ℓ, then the point where the line intersects the θ_ℓ-axis is given by $-d_i/a_\ell$. This is usually called the intercept of the line with that axis. For that reason, d is usually called the *intercept parameter*.

The d-parameter is not a difficulty parameter in the usual sense of a UIRT model because it does not give a unique indicator of the difficulty of the item. Instead, the negative of the intercept term divided by an element of the discrimination parameter vector gives the relative difficulty of the item related to the corresponding coordinate dimension. For example, Fig. 4.4 shows the .5 line for the item shown in Fig. 4.2 as line AC extended. The intersection of the line AC with the θ_1 coordinate axis is $-d/a_1 = -(-.7)/.5 = 1.4$. This is the location of C on the graph. Similarly, the location of A on the θ_2 axis is $-(-.7)/1.5 = .47$. This indicates that if θ_2 were 0,

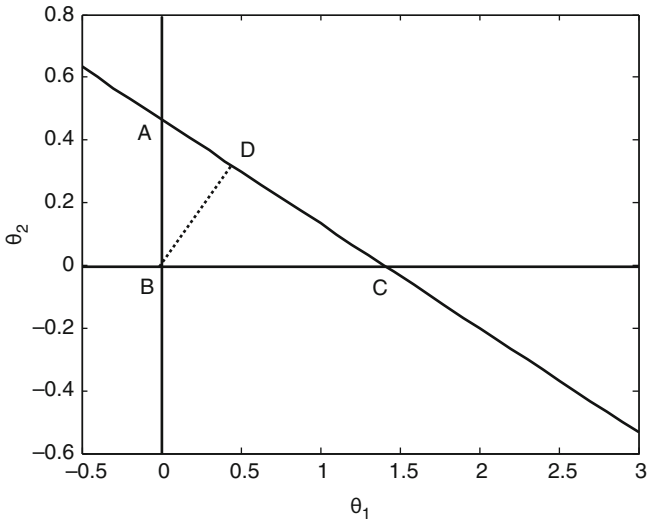

Fig. 4.4 Relationship of the .5 probability line with the coordinate axes for an item with $a_1 = .5$, $a_2 = 1.5$, and $d = -.7$

then a person would need a value of θ_1 of 1.4 to have a probability of .5 of a correct response. But, if θ_1 were 0, then a person would need only .47 on the θ_2 dimension to have a .5 probability of correct response. The conditional nature of these statements makes the actual intercepts awkward indicators of the item difficulty. A useful alternative is the distance of the line from the origin of the space. This is indicated by the dotted line BD on the figure.

The distance between the origin and the line can be derived from the properties of similar triangles. The complete derivation is left to the reader. If the length of BD is represented by b, the general expression for the distance of the line from the origin is given by

$$b = \frac{-d}{\sqrt{\mathbf{a}\mathbf{a}'}} = \frac{-d}{\sqrt{\sum_{v=1}^{m} a_v^2}}. \tag{4.9}$$

The value of b has the same interpretation as the b-parameter for UIRT models. To avoid confusion with the UIRT models, b is often called MDIFF in the MIRT literature. When MDIFF is 0, the .5 contour line goes through the origin of the space. This means that one of the vectors of $\boldsymbol{\theta}$-parameters that will yield a .5 probability of response is the 0-vector (a vector with all m elements equal to 0). Of course, any point on the .5 probability line also yields the same probability of response, so there are many other $\boldsymbol{\theta}$-vectors that will yield a .5 probability.

There are many variations on the M2PL model that have very similar properties. They all have equiprobable contours that are straight lines. They are also

compensatory in the sense described above. Several variations and extensions of the M2PL model are described in the next sections of this chapter.

Multidimensional extension of the three-parameter logistic model. A fairly straightforward extension of the M2PL model provides for the possibility of a non-zero lower asymptote to the model. This is a multidimensional extension of the three-parameter logistic UIRT model described in Chap. 2. This model is typically labeled the M3PL model. The mathematical expression for the model is given in (4.10) using the symbols as previously defined.

$$P(U_{ij} = 1|\boldsymbol{\theta}_\mathbf{j}, \mathbf{a_i}, c_i, d_i) = c_i + (1 - c_i)\frac{e^{\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i}}{1 + e^{\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i}}. \tag{4.10}$$

The M3PL model was designed to account for observed empirical data such as that provided in Lord (1980), which shows that examinees with low capabilities have a nonzero probability of responding correctly to multiple-choice items. Because the process of selecting a correct response for individuals with low capabilities does not seem to be related to the constructs assessed by the test item, the model contains a single lower asymptote, or pseudo-guessing, parameter, c_i, to specify the probability of correct response for examinees with very low values in $\boldsymbol{\theta}$.

The item response surface for item response data that can be modeled with two coordinate dimensions is presented in Fig. 4.5. The graph of the model shows that the lines of equal probability are straight lines as was the case for the M2PL model and the form of the surface is basically the same. The major difference is that the surface asymptotes to c_i rather than continuing down to a probability of response of 0.

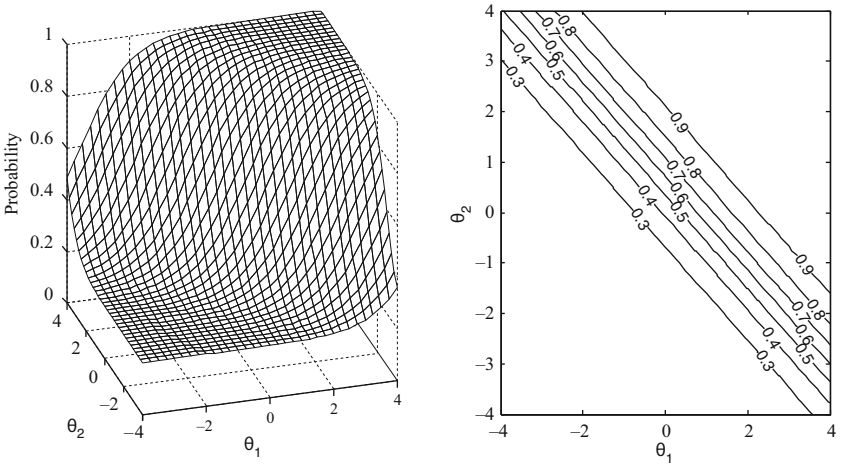

Fig. 4.5 Surface plot and contour plot for probability of correct response for an item with $a_1 = 1.3$, $a_2 = 1.4$, $d = -1$, $c = .2$

Multidimensional extension of the Rasch model. It is tempting to consider the multidimensional extension of the Rasch model as simply the M2PL model with all of the a-parameters set to 1.0. This would be equivalent to the relationship between the Rasch one-parameter logistic model and the two-parameter logistic model for the UIRT case. However, an analysis of the consequences of setting all of the a-parameters in the M2PL model to 1.0 shows that this type of generalization of the UIRT case does not give a useful result.

The general form of the exponent of the M2PL model is given by

$$a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \ldots + a_{im}\theta_{jm} + d_i. \tag{4.11}$$

If all of the a-parameters are set equal to the same value, say a_{i*}, they can be factored from the expression in (4.11) to yield

$$a_{i*}(\theta_{j1} + \theta_{j2} + \ldots + \theta_{jm}) + d_i. \tag{4.12}$$

For Person j at any given moment, the θ-coordinates are fixed and the sum of them takes on a particular value that can be represented as θ_*. Substituting the sum of the θ-coordinates into (4.12) yields $a_{i*}\theta_{j*} + d_j$. Finally, if $b_i = -d_i/a_{i*}$, the exponent can be represented as

$$a_{i*}(\theta_{j*} - b_i). \tag{4.13}$$

If all of the a_{i*} values are set to one, the exponent has the same form as the simple UIRT Rasch model. The only difference is that the ability parameter is a value that is the sum of coordinates rather than what is usually interpreted to be the level on a single construct. Trying to create a multidimensional version of the Rasch model by setting the a-parameters to 1.0 only yields a UIRT version of the Rasch model with a more complex definition for the person parameter.

The multidimensional generalization of the Rasch model is more complex. The approach that currently appears in the psychometric literature (Adams et al. 1997) is an adaptation of the general Rasch model presented in Chap. 3 (3.8). The model as specified in Adams et al. (1997) is for the general case that includes both dichotomously and polytomously scored test items. The general form of the model is presented first using the notation from the original article. Then the simpler case for dichotomously scored items is presented using the same notation as the other models presented in this book.

The expression for the full model is presented in (4.14). This model is for a test item that has the highest score category for Item i equal to K_i. The lowest score category is 0. This implies that the number of score categories is $K_i + 1$. For the dichotomous case, $K_i = 1$ and there are two score categories, 0 and 1. The score category is represented by k. The random variable X_{ik} is an indicator variable that indicates whether or not the observed response is equal to k on Item i. If the score is k, the indicator variable is assigned a 1 – otherwise, it is 0. For the dichotomous case, if $X_{i1} = 1$, the response to the item was a correct response and it was assigned a score of 1.

$$P(X_{ik} = 1|\mathbf{A}, \mathbf{B}, \boldsymbol{\xi}, \boldsymbol{\theta}) = \frac{e^{\mathbf{b_{ik}}\boldsymbol{\theta} + \mathbf{a_{ik}}'\boldsymbol{\xi}}}{\sum_{k=0}^{K_i} e^{\mathbf{b_{ik}}\boldsymbol{\theta} + \mathbf{a_{ik}}'\boldsymbol{\xi}}}, \tag{4.14}$$

where $\mathbf{A}$ is a design matrix with vector elements $\mathbf{a_{ik}}$ that select the appropriate item parameter for scoring the item; $\mathbf{B}$ is a scoring matrix with vector elements $\mathbf{b_{ik}}$ that indicate the dimension or dimensions that are required to obtain the score of k on the item; $\boldsymbol{\xi}$ is a vector of item difficulty parameters; and $\boldsymbol{\theta}$ is a vector of coordinates for locating a person in the construct space.

For the dichotomous case, using the same notation as the other models, the multidimensional Rasch model is given by

$$P(U_{ij} = 1|\mathbf{a_i}, d_i, \boldsymbol{\theta_j}) = \frac{e^{\mathbf{a_i}\boldsymbol{\theta_j}' + d_i}}{1 + e^{\mathbf{a_i}\boldsymbol{\theta_j}' + d_i}}, \tag{4.15}$$

where $\mathbf{a_i}$ is a vector such that $\mathbf{a_i} = \mathbf{b_{ik}}$ and d_i is a scalar value equal to $\mathbf{a}'_{\mathbf{ik}}\boldsymbol{\xi}$. Note that when $k = 0$ in (4.14) the exponent of e is 0 so that term of the sum in the denominator is equal to 1.

Equation (4.15) and (4.5) appear to be identical. The difference between the two is the way that the $\mathbf{a}_i$ vector is specified. In (4.5), $\mathbf{a}_i$ is a characteristic of Item i that is estimated from the data. In (4.15), $\mathbf{a}_i$ is a characteristic of Item i that is specified by the test developer. In the case of the model in (4.5), statistical estimation procedures are used to determine the elements of $\mathbf{a}_i$ that will maximize some criterion for model/data fit. Except for the usual monotonicity constraint that requires the values of the elements of $\mathbf{a}_i$ be positive, the elements can take on any values. For the model in (4.15), the values are specified by the analyst and they typically take on integer values. Adams et al. (1997) specified two variations for the model – between item and within item dimensionality. For between item dimensionality, the $\mathbf{a}_i$-vector has elements that are all zeros except for one element that specifies the coordinate dimension that is measurement target for the item. That is, the test developer specifies the dimension that is the target for the item. In a sense, the test developer estimates the elements of the $\mathbf{a}_i$-vector rather than obtaining estimates through the usual statistical estimation procedures.

For within item dimensionality, the $\mathbf{a}_i$-vector has more than one nonzero element. The test developer can indicate that performance on the test item is influenced by more than one of the coordinate dimensions. For the two-dimensional case, $\mathbf{a}_i$-vectors of [1 0] or [0 1] would indicate between item dimensionality. The first vector would specify that the item was only affected by level on coordinate dimension 1 and the second vector specifies that the item is only affected by level on coordinate dimension 2. A specification for within item dimensionality might have a vector such as [1 1] indicating that the item is affected equally by both coordinate dimensions. Other alternatives such as [1 2] or [3 1] are possible. The quality of the fit of the model to the data will depend on how well the test developer specifies the values of the $\mathbf{a}_i$-vector.

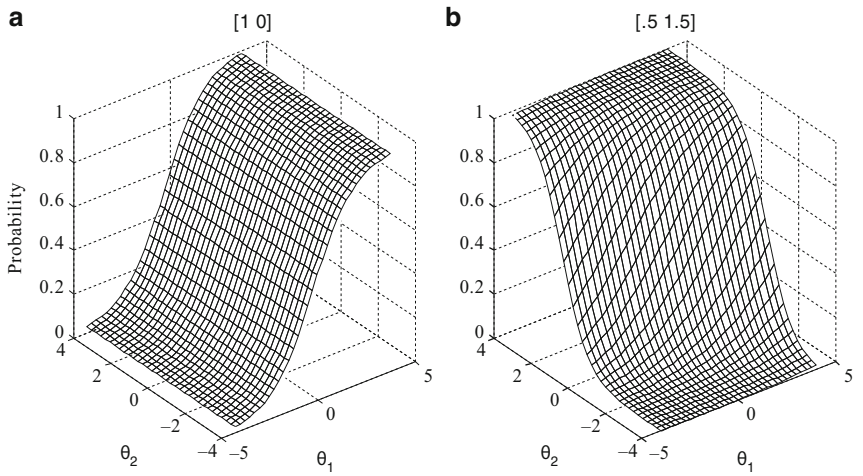

Fig. 4.6 Two-dimensional Rasch model surfaces with $\mathbf{a}_i = [1\ 0]$ and $[.5\ 1.5]$ and $d_i = 0$

The reason for specifying the $\mathbf{a}_i$-vector rather than using statistical estimation methods to determine the values is that the model then has observable sufficient statistics for the person and item parameters (Rasch 1962). The sufficient statistic for the $\boldsymbol{\theta}$-vector for a person is the sum of the $\mathbf{a}_i$-vectors for the items a person answers correctly. The sufficient statistic that is related to the d_i-parameter is the sum over people of the $\mathbf{a}_i$-vectors for the correct responses to Item i. This is simply the number of times Item i is answered correctly times $\mathbf{a}_i$.

The form of the item response surface for the multidimensional Rasch model for dichotomously scored items is the same as that shown in Fig. 4.3, but the orientation of the surface is dependent on the specification of values in the $\mathbf{a}_i$-vector. Figure 4.6 shows surfaces for the two-dimensional case with $\mathbf{a}_i$-vectors [1 0] and [.5 1.5] and $d_i = 0$ for both examples. The surface in panel a does not differentiate at all in probability for differences along θ_2 – the item only measures θ_1. The item represented in panel b has changes in probability for changes in θ_2 and slight changes along θ_1.

The model given in (4.14) has many other variations. Some of those will be described later in this chapter when multidimensional models for polytomously scored items are presented. The reader is referred to Adams et al. (1997) for a more detailed description of other variations of the model.

Multidimensional extension of the normal ogive model. As indicated in Chap. 3, much of the original work on MIRT was done using the normal ogive form to represent the relationship between the location in the multidimensional space and the probability of a correct response to a test item. The normal ogive form is still used as the basis of statistical estimation programs. These programs and the underlying statistical methods are described in Chap. 6.

The general form for the multidimensional extension of the normal ogive model (Bock and Schilling 2003; McDonald 1999; Samejima 1974) is given by

$$P(U_{ij} = 1|\boldsymbol{\theta}_\mathbf{j}, \mathbf{a_i}, c_i, d_i) = c_i + (1 - c_i)\frac{1}{\sqrt{2\pi}} \int_{-z_i(\boldsymbol{\theta}_\mathbf{j})}^{\infty} \mathrm{e}^{\frac{-t^2}{2}} \mathrm{d}t, \tag{4.16}$$

where $z_i(\boldsymbol{\theta}_\mathbf{j}) = \mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i$ and the other symbols have been previously defined. If $c_i = 0$, the result is the normal ogive version of the multidimensional two-parameter logistic model. That form of the model defines the probability of correct response for an item as the area under a standard normal distribution from $-z_i(\boldsymbol{\theta}_\mathbf{j})$ to infinity. Because of the symmetry of the normal distribution, this is the same as the area below $z_i(\boldsymbol{\theta}_\mathbf{j})$.

The form of the surface defined by (4.16) is essentially the same as that defined by (4.10). Camilli (1994) summarized the work on comparing the normal ogive and logistic functions. He included the mathematical proof by Haley (1952) showing that the normal distribution function and the logistic function differ by less than .01 in probability when the constant 1.702 is included in the exponent of the logistic function. More explicitly, for z as defined above,

$$|\Phi(z) - \Psi(1.702z)| < .01 \quad \text{for } -\infty < z < \infty, \tag{4.17}$$

where Φ is the cumulative normal ogive function and Ψ is the cumulative logistic function.

For the parameters in (4.10) and (4.16) to have essentially the same meaning, the exponent of (4.10) has to be changed to $1.702(\mathbf{a_i}\boldsymbol{\theta}_\mathbf{j}' + d_i)$. Multiplying by the constant 1.702 changes the scale for the parameters in the logistic model, but has no other effect on the form of the surface. Figure 4.7 shows examples of the surfaces for

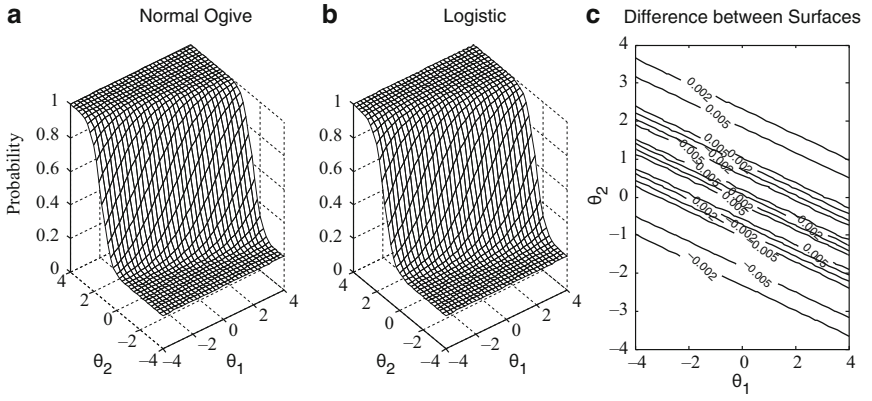

Fig. 4.7 Normal ogive and logistic surfaces, and difference between the surfaces for a test item with $a_1 = .5$, $a_2 = 1.5$, $d = 0$, $c = .2$. Note that $D = 1.702$ is included in the exponent of the logistic model

an item that can be modeled in two-dimensions using both the logistic and normal ogive models. The difference in the two surfaces is also shown.

The surfaces shown in panels a and b are virtually indistinguishable. The contour plot of the differences between the two surfaces has less than .01 in all cases, and the differences were the same along the lines of equal probability. This indicates that the surfaces have the same orientation to the coordinate axes – the lines of equal probability are the same. Because of the close similarity of the surfaces defined by the two models, the parameters estimated from the models are often used interchangeably as long as the 1.702 is included in the exponent of the logistic model.

All of variations of the logistic model presented in this chapter can also be implemented for the normal ogive model by modifying (4.16). If $c_i = 0$, the result is the multivariate generalization of the two-parameter normal ogive model. The z-function can also be defined with prespecified $\mathbf{a}_\mathrm{i}$ vectors to yield a confirmatory model that is like the multivariate version of the Rasch model.

4.1.1.2 Partially Compensatory Extensions of UIRT Models

One of the continuing theoretical issues raised about the model presented in (4.5) is the compensatory nature of the model. If any $m - 1$ coordinates for a person are very low, the person can still have a very high probability of correct response if the mth coordinate is sufficiently high. Sympson (1978) argued that this hypothesized compensation is not realistic for some types of test items. He presents an example of a mathematics test item that requires both arithmetic computation skills and reading skills. His example has similar characteristics to the test item presented at the beginning of this chapter. Sympson (1978) hypothesizes that a person with reading skill that is very low will not be able to determine the problem that needs to be solved. The contention is that even if such a person had very high mathematics computation skills, they would not be able to determine the correct answer because of lack of understanding of the problem. To model the situation he considered, Sympson (1978) proposed the following expression for the interaction between the person and the test item, where all of the symbols have the same meaning as in previous equations.

$$P(U_{ij} = 1|\boldsymbol{\theta}_\mathbf{j}, \mathbf{a}_\mathbf{i}, \mathbf{b}_\mathbf{i}, c_i) = c_i + (1 - c_i)\left(\prod_{\ell=1}^{m} \frac{\mathrm{e}^{1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}}{1 + \mathrm{e}^{1.7a_{i\ell}(\theta_{j\ell} - b_{i\ell})}}\right). \tag{4.18}$$

The expression in (4.18) has two main parts. The term on the far right is the product of terms that have the form of the 2PL UIRT model that was described in Chap. 2. In a sense, each of these terms gives the probability of being successful on one component of the item – for example, the reading or the mathematics components of the example item. These components of the test item are considered as independent activities so that the probability of doing all of them correctly is the product of the probabilities of doing each part correctly. The other part of the expression to the right of the equal sign provides a nonzero lower asymptote for the

model, c_i. This part of the model has the same function as in (4.10). There is only one c_i-parameter for each item. Sympson (1978) did not believe that there was a lower asymptote for each task in the item, but only for the item overall.

The form of the surface defined by (4.18) can be investigated by considering the case when $c_i = 0$ and the probability of correct response to the test item is some constant value, k. Further, if the 2PL terms in the product are represented by p_ℓ, where ℓ is the dimension of interest, the simplified version of the model becomes

$$k = \prod_{\ell=1}^{m} p_\ell. \tag{4.19}$$

For the simple case of only two dimensions, the expression is simply $k = p_1 p_2$. This is the equation for a hyperbola with values in the probability metric. The hyperbolas defined by this function are shown in the left panel of Fig. 4.8 for $k = .25, .50,$ and .75. The hyperbolas do not continue on to the asymptotic values because the probabilities are constrained to the range from 0 to 1. As a result, only segments of the hyperbolas are shown.

The pairs of probabilities that are specified by each point along one of the parabolas can be transformed to the θ scale through the item characteristic curve for the item. This process results in a pair of θ coordinates for each point on a hyperbola. These points can be plotted in the θ-space to show the sets of coordinates that yield the specified probability of correct response to a test item that is well modeled by (4.18). Note that for a specified probability of correct response to a test item, only one hyperbola is defined for a specific number of dimensions. If $k = .5$ and the item is modeled with two dimensions, the corresponding hyperbola is the one shown by the dashed line in Fig. 4.8. The θ-vectors that correspond to the probability pairs represented by the hyperbola are dependent on the item parameters for the item. For

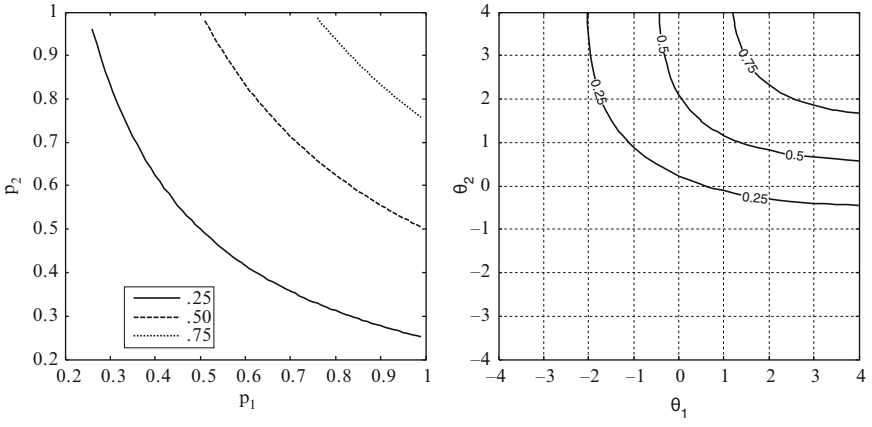

Fig. 4.8 Curves of equal probability of correct response for the noncompensatory model with two coordinates and $c_i = 0$

example, if the item parameters for an item modeled in two dimensions are $c_i = 0$, $a_{i1} = .7$, $a_{i2} = 1.1$, $b_{i1} = -.5$, and $b_{i2} = .5$, the curves in the θ-space that correspond to the hyperbolas in the left panel of Fig. 4.8 are shown in the right panel of the same figure.

The curves in the right panel of Fig. 4.8 are not true hyperbolas, although they do asymptote to specific values. The .5 curve asymptotes to the values of the b-parameters, -.5 and .5. This is a result of the partially compensatory nature of the model. If θ_1 is equal to -.5, then the probability of correctly responding to the first component of the item is .5. The actual probability of correct response to the item as a whole depends on the probability of the second component. For example, if θ_2 were equal to .5, then the probability of responding correctly to the second component of the item would be .5, and the overall probability of correct response would be only .25, that is, the product $p_1 p_2$. As θ_2 increases, the probability of correct response increases. But, even if θ_2 is positive infinity yielding a probability of 1 for the second component, the overall probability of correct response would only be .5. Thus, the probability of correct response for an item that follows this model can never be greater than the probability for the component with the lowest probability.

The item response surface for the item with the parameters given above and $c_i = .2$ is shown in Fig. 4.9. Inspection of the figure will show that the surface has a lower asymptote to the value of the c-parameter. Also, the curvature of the surface shows the partially compensatory nature of the model. The probability of correct response for low values of either θ_1 or θ_2 or both is close to the lower asymptote

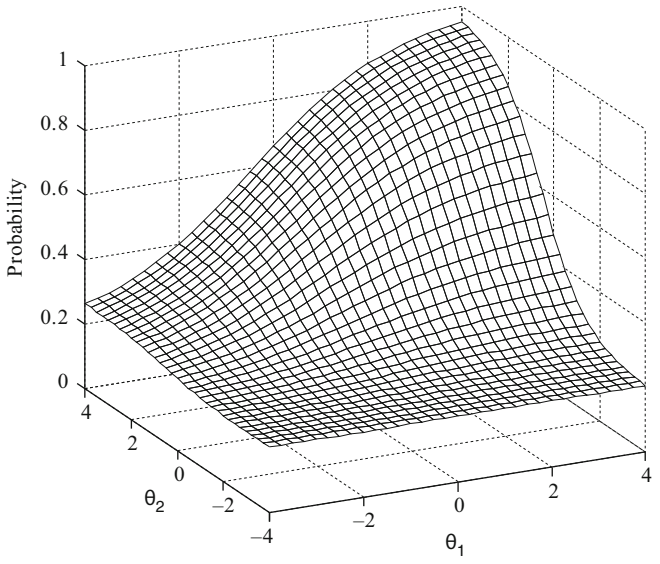

Fig. 4.9 Item response surface for the partially compensatory model when $a_1 = .7$, $a_2 = 1.1$, $b_1 = -.5$, $b_2 = .5$, and $c = .2$

value. Only when both θ-values are high does the model yield a high probability of correct response.

Another unique feature of this model is that the interpretation of the item parameters changes as the number of dimensions increase. Those familiar with UIRT probably know that when θ is equal to the difficulty parameter for the item, b, the probability of correct response to the test item is .5 if there is no lower asymptote parameter in the model. A simple example is if $\theta = 0$ and $b = 0$, the probability of correct response is .5. The same is true for the compensatory model. If the $\boldsymbol{\theta}$-vector is 0 and the d-parameter is 0, the probability of correct response to the item is .5 when there is no lower asymptote parameter. However, that is not the case for the partially compensatory model. If the $\boldsymbol{\theta}$-vector and the $\mathbf{b}$-vector are both 0 for the two-dimensional case, the probability of correct response is .25. For the three dimensional case, it is .125. In general, for the case when both the $\boldsymbol{\theta}$ and $\mathbf{b}$-vectors are 0, the probability of correct response is $.5^m$, where m is the number of dimensions used in the model.

The value of a common b-parameter that yields a .5 probability of correct response depends on the corresponding a-parameter. However, it is useful to consider the special case when $\boldsymbol{\theta}$ is the zero vector, all a-parameters are .588 (note that the constant 1.7 is in the model), and the c-parameter is 0. If all the b-parameters are the same, what must they be for the probability of correct response for the test item to equal .5 for different numbers of dimensions? To get the answer, the left side of (4.18) is set to .5 and the equation is solved for $\mathbf{b}$ under the above constraints. The results are provided in Table 4.3.

The results in Table 4.3 show that as the number of dimensions increases, the value of the b-parameter must be reduced to maintain the same probability of correct response. This is a direct result of the fact that the model contains a product of the probabilities of success on each component. Whenever the probability of success is less than 1.0, adding another component results in multiplying by a number less than 1.0. This reduces the overall probability of a correct response. To compensate for this, the components must be easier as reflected by the reduced magnitude of the b-parameters.

The model presented in (4.18) is a multidimensional extension of the three-parameter logistic UIRT model. Whitely[1] (1980b) suggested using a simplified version of (4.18) to model the cognitive processes in test items. Maris (1995) also

Table 4.3 b-parameter required for a probability of correct response of .5 for different numbers of dimensions

Number of dimensions	b-parameter
1	0
2	−.88
3	−1.35
4	−1.66
5	−1.91
6	−2.10

[1] Whitely now publishes under the name Embretson

suggested this model and labeled it the conjunctive Rasch model. The model is an extension of the one-parameter logistic model. This model is presented in (4.20). Although the terms in the product are equivalent to the Rasch UIRT model, the multidimensional model does not match the requirements for a Rasch model because there is not an observable sufficient statistic for the person parameter vector. Adams et al. (1997) indicate that the model in (4.20) can be considered as a Rasch model if it is considered as an item with 2^m possible response categories. A vector of 0 s and 1 s is developed to describe the success on each cognitive component of the model, and each of the possible vectors is considered as a response to the full item. This approach is based on the assumption that success on the various cognitive components is independent. Whitely (1980b) also suggested a model of the same general form that assumed that success on component ℓ was dependent on success on component $\ell - 1$.

$$P(U_{ij} = 1|\boldsymbol{\theta}_{\mathbf{j}}, \mathbf{b_i}) = \prod_{k=1}^{m} \frac{e^{(\theta_{jk} - b_{ik})}}{1 + e^{(\theta_{jk} - b_{jk})}}. \tag{4.20}$$

4.1.1.3 Comparison of Compensatory and Partially Compensatory Models

The models that combine the effect of multiple θs using a linear function in the exponent of the model (compensatory models) as in (4.5) and the ones that combine the θs through a product of terms (partially compensatory models) as in (4.18) are quite different in their philosophical underpinnings and in mathematical form. Some researchers (e.g., Maris 1995) may prefer one model over the other because it better matches hypotheses about how persons interact with the test items. The partially compensatory product models are consistent with the hypothesis that test items have different parts related to different skills or knowledge and that overall success requires success on each part. The compensatory models are more consistent with a more holistic view of the interaction of persons and test items. Persons bring all of their skills and knowledge to bear on all aspects of the items. Ultimately, the usefulness of the models will be determined by how accurately they represent the responses from actual test items. Only a few studies were available at the time this book was written that compared the fit of the two types of models to the same data. One study that was identified, Bolt and Lall (2003), found that the compensatory model fit item response data from an English usage test better than the partially compensatory model in (4.20). Note that this model does have all of the a-parameters set equal to 1. That study also compared the models on the fit to data generated from the other model. The compensatory model fit partially compensatory data almost as well as the partially compensatory model. The partially compensatory model did not fit the compensatory data very well.

The results reported by Spray et al. (1990) may provide some insight into the Bolt and Lall (2003) results. They carefully selected parameters for the compensatory model in (4.15) so that generated item response data would have the same characteristics such as internal consistency reliability, item difficulty distribution, and item discrimination as real test data. They then generated 2,000 $\boldsymbol{\theta}$-vectors and

Table 4.4 Item parameters for the partially compensatory and compensatory models for the same test item

Partially compensatory model		Compensatory model	
Parameters	Values	Parameters	Values
a_1	1.26	a_1	.90
a_2	1.60	a_2	1.31
b_1	−.92	d	−.67
b_2	−.15		

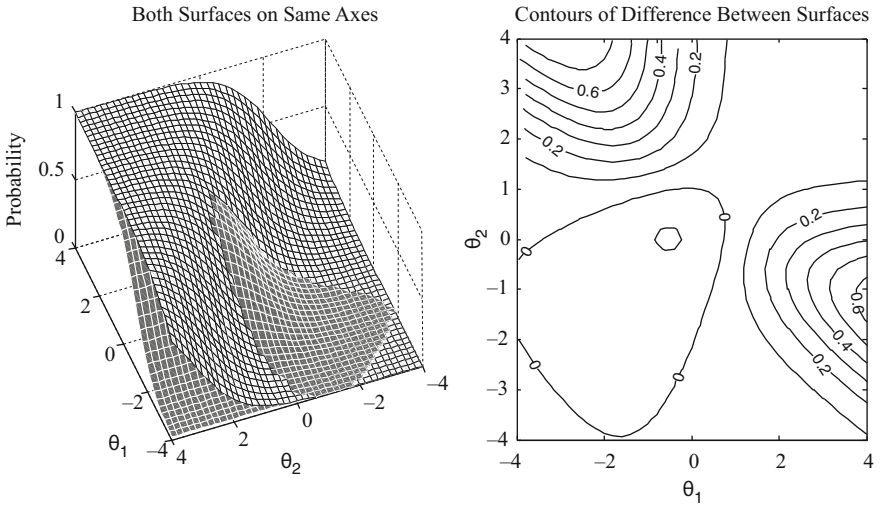

Fig. 4.10 Comparison of partially compensatory and compensatory item response surfaces

computed the probability of correct response for each vector and the compensatory parameters for each item. Then, assuming $\boldsymbol{\theta}$-vectors known, the parameters of the partially compensatory model were estimated that minimized the squared differences in probabilities of correct response. The results were item response surfaces that gave the same p-values for each item assuming a standard bivariate normal distribution of $\boldsymbol{\theta}$.

Table 4.4 gives the item parameters for the two models that gave best match to the item response surfaces. The c_i parameter was set to zero for the item. Figure 4.10 shows the differences in the representation of the interaction between persons and the test item in two different ways. The left panel of Fig. 4.10 shows the item response surfaces for the models on the same set of axes. The dark gray surface is for the partially compensatory model and the white surface is for the compensatory model. The graph shows that the two surfaces intersect, with higher probabilities for the partially compensatory surface $\boldsymbol{\theta}$-vectors when both elements are below 0.

The right panel in Fig. 4.10 shows a contour plot for the differences between the two surfaces. The enclosed curve labeled "0" represents the intersection of the two

surfaces. The probability of correct response is the same for the two models for $\boldsymbol{\theta}$-vectors that fall on that curve. The other curves are labeled with the difference in probability of correct response for the two surfaces. Those contours show that the models give quite different probabilities of correct response for $\boldsymbol{\theta}$-vectors near $(-3, 4)$ and $(4, -1)$. In both of these regions of the $\boldsymbol{\theta}$-space, the compensatory model gives probabilities of correct response that are over .6 higher than the partially compensatory model.

The two comparisons of the item response surfaces show that for $\boldsymbol{\theta}$vectors along the diagonal from $(-4, -4)$ to $(4, 4)$ there is little difference in the probability of correct response for the two models. For real tests with θ-elements that may be positively correlated, the two models may give very similar results. It is only when the elements of the $\boldsymbol{\theta}$-vector are very different that the models predict different rates of responding correctly to the test item.

The results for the test analyzed by Bolt and Lall (2003) suggest that the $\boldsymbol{\theta}$-coordinates tend to function in a compensatory way for the items on the test. However, for other tests, the partially compensatory model may be more appropriate. Ultimately, empirical studies of the usefulness of the models for representing the item response data are needed to determine which form of the model more accurately represents the interactions between persons and test items.

4.1.2 MIRT Models for Test Items with More Than Two Score Categories

Although test items that are scored using more than two score categories have been used for a long time, the development of item response theory models for these item types is a relatively new development. Chapter 2 provides a summary of some of these IRT models under the category polytomous IRT models (see Sect. 2.3). The polytomous IRT models have been extended to allow the person characteristics to be represented by $\boldsymbol{\theta}$-vectors. Muraki and Carlson (1993) produced an extension of the graded response model, and there have been recent extensions of the generalized partial credit model (Yao and Schwarz 2006). The following sections of this chapter describe the extensions of the generalized partial credit, partial credit, and graded response models. These models all fall under the label of compensatory models. At the time of the writing of this book, no partially compensatory polytomous models have been proposed.

4.1.2.1 Multidimensional Generalized Partial Credit Model

The multidimensional extension of the generalized partial credit (MGPC) model is designed to describe the interaction of persons with items that are scored with more than two categories. The maximum score for Item i is represented by K_i. To be consistent with the way dichotomous items are scored, the lowest score is assumed

to be 0 and there are $K_i + 1$ score categories overall. The score assigned to a person on the item is represented by $k = 0, 1, \ldots, K_i$. The mathematical representation of the MGPC model is given by the following equation,

$$P(u_{ij} = k|\boldsymbol{\theta}_j) = \frac{e^{k\mathbf{a_i}\boldsymbol{\theta}'_\mathbf{j} - \sum_{u=0}^{k} \beta_{iu}}}{\sum_{v=0}^{K_i} e^{v\mathbf{a_i}\boldsymbol{\theta}'_\mathbf{j} - \sum_{u=0}^{v} \beta_{iu}}}, \tag{4.21}$$

where β_{iu} is the threshold parameter for score category u, β_{i0} is defined to be 0, and all other symbols have there previously defined meaning. The representation of the model given here is a slight variation of the form given in Yao and Schwarz (2006).

There are two important differences between the equation for the MGPC model and that for the GPC model given in (3.33). First, the model does not include separate difficulty and threshold parameters. Second, because $\boldsymbol{\theta}$ is a vector and the βs are scalars, it is not possible to subtract the threshold parameter from $\boldsymbol{\theta}$. Instead, the slope/intercept form of the generalized partial credit model is used as the basis of the multidimensional generalization, $a\theta + d$, but with the sign of the intercept reversed. The result is that the βs cannot be interpreted in the same way as the threshold parameters in the UIRT version of the model. This will be discussed in more detail after presenting the form of the item response surface.

The item response surfaces for the MGPC model for the case when the item/person interaction can be represented in a space with two coordinate dimensions is presented in Fig. 4.11. The test item represented here has scores from 0 to 3. The item parameters for the model are $\mathbf{a_i} = [1.2\ .7]$ and $\beta_{iu} = 0, -2.5, -1.5, .5$.

Figure 4.11 presents four surfaces – one for each possible item score. The darkest surface to the left is for the score of 0. The probability of that score decreases as the

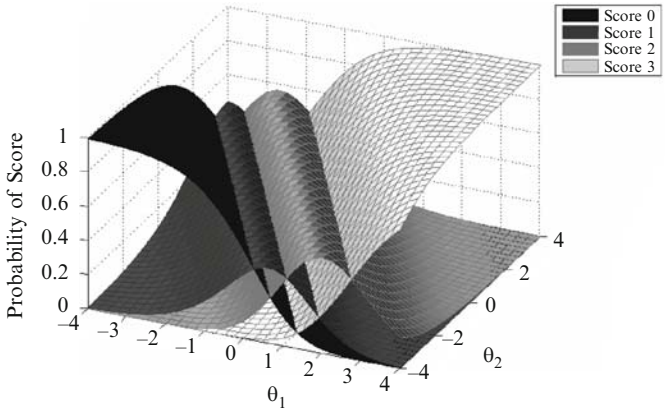

Fig. 4.11 Item response surfaces for MGPC with item parameters $\beta_{iu} = 0, -2.5, -1.5, .5$, and $\mathbf{a_i} = [1.2\ .7]$

θ-coordinate increases on either dimension. The surfaces for scores 1 and 2 first increase and then decrease as the θ-coordinates increase. The surface for the score of 3 increases with an increase in any of the θ-coordinates. The surfaces for the scores of 0 and 3 have upper asymptotes of 1 and lower asymptotes of 0. The other two surfaces have ridge-shaped forms with lower asymptotes of 0.

The intersections between the surfaces for adjacent score categories are over a straight line in the $\boldsymbol{\theta}$-plane. In general, the line is the set of points in the $\boldsymbol{\theta}$-plane where the probabilities of obtaining the adjacent scores are equal. This set of points can be obtained by finding the solution to (4.22). This equation is specified by setting the exponents in the numerator of (4.21) for adjacent score categories equal to each other.

$$k\mathbf{a_i}\boldsymbol{\theta_j}' - \sum_{u=0}^{k} \boldsymbol{\beta}_{iu} = (k+1)\mathbf{a_i}\boldsymbol{\theta_j}' - \sum_{u=0}^{k+1} \boldsymbol{\beta}_{iu}. \tag{4.22}$$

Some algebraic manipulation results in the following solution for the intersection between the kth and $(k+1)$th surfaces. This is the equation for a line in the m-dimensional space used to represent the item. Note that the only part of this expression that changes for different adjacent score categories is the intercept term, β. As was the case for the UIRT version of the generalized partial credit model, this parameter controls the location of the thresholds between score categories.

$$0 = \mathbf{a_i}\boldsymbol{\theta_j}' - \boldsymbol{\beta}_{i,k+1}, \quad k = 0, \ldots, k-1. \tag{4.23}$$

Equation (4.21) gives the expression for the probability of each score for a test item. A surface that defines the expected score on the test item for a person with a particular $\boldsymbol{\theta}$-vector is given by

$$\boldsymbol{E}(u_{ij}|\boldsymbol{\theta_j}) = \sum_{k=0}^{K_i} \boldsymbol{k}\boldsymbol{P}(u_{ij} = \boldsymbol{k}|\boldsymbol{\theta_j}). \tag{4.24}$$

The surface defined by (4.24) for the item shown in Fig. 4.11 is given in Fig. 4.12.

This surface has the appearance of the item response surface for a dichotomously scored item for a compensatory model, but the upper asymptote for the surface is 3, the maximum score on the item. The MGPC is a compensatory model in the same sense as the UIRT version of the model in that a high value on θ_v can compensate for a low value on θ_w resulting in a high expected score on the item. This effect can be seen in the figure for the point (4, −4) that has an expected score near 3.

4.1.2.2 Multidimensional Partial Credit Model

There are a number of simplifications of the multidimensional version of the generalized partial credit model that have the special properties of the Rasch model. That is, they have observable sufficient statistics for the item- and person-parameters. Kelderman and Rijkes (1994) present the general form for one multidimensional extension of the Rasch model to the polytomous test item case. Their model is

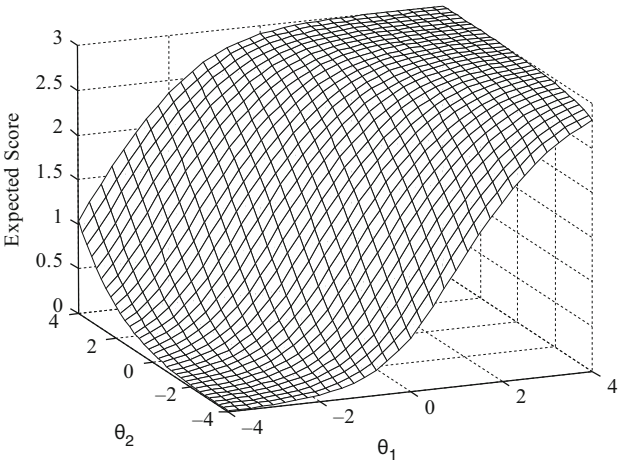

Fig. 4.12 Expected score surface for the item in Fig. 4.11

presented below (4.25) using slightly different symbols than their original presentation to facilitate comparisons with the other models presented in this book. A very similar model is presented by Adams et al. (1997).

$$\boldsymbol{P}(u_{ij} = \boldsymbol{k}|\boldsymbol{\theta}_{\mathbf{j}}) = \frac{e^{\sum_{\ell=1}^{m}(\theta_{j\ell}-b_{i\ell k})W_{i\ell k}}}{\sum_{r=0}^{K_i} e^{\sum_{\ell=1}^{m}(\theta_{j\ell}-b_{i\ell r})W_{i\ell r}}}, \tag{4.25}$$

where $b_{i\ell k}$ is the difficult parameter for Item i on dimension ℓ for score category k, and $W_{i\ell k}$ is a predefined scoring weight for Item i related to dimension ℓ and score category k. The other symbols have the same meaning as in previous equations.

The key to the functioning of this model is the specification of the matrix of weights, $W_{i\ell k}$. Suppose that a test item has $K_i = 3$ score categories: 0, 1, and 2. Also assume that the item is sensitive to differences on two dimensions. The weight matrix for such an item might be specified as $\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$. In this matrix, the rows represent the score categories, k, and the columns represent the dimensions, ℓ. This matrix indicates that a 0 response to the item indicates a 0 score on both dimensions, a 1 response indicates that the part of the item related to Dimension 1 was correct but that on Dimension 2 was incorrect, and a score of 2 indicated that the examinee successfully performed the components of the item that are related to both of the dimensions.

For this special case, when $k = 0$, the term in the numerator of (4.25) is $e^0 = 1$. For the other values of k, (4.25) simplifies to

$$\boldsymbol{P}(u_{ij} = k|\boldsymbol{\theta}_{\mathbf{j}}) = \frac{e^{\sum_{\ell=1}^{k}(\theta_{j\ell} - b_{i\ell k})}}{1 + \sum_{r=1}^{K_i} e^{(\theta_{j\ell} - b_{i\ell r})}}, \quad k = 1, \ldots, K_i. \tag{4.26}$$

The developers of this model note that there is an indeterminacy in the estimation of the $b_{i\ell k}$ parameters so they set the parameters equal across the response categories, $k = 1, 2, \ldots, K_i$. The items have different difficulty parameters for the different dimensions, but the same for response categories within a dimension. This means that the item functions as a series of dichotomous items for each dimension with a difficulty for that dimension.

The item response surfaces for each score category for a test item scored 0, 1, or 2 using this model are shown in Fig. 4.13. This is for the two-dimensional case with the scoring matrix given on the previous page. The $b_{i\ell k}$ parameters for the item are -1 for dimension 1 and $+1$ for dimension 2. Careful study of Fig. 4.13 will show that the surface for a score of zero is highest for $\boldsymbol{\theta}$-vector $(-4, -4)$ and lowest for the vector $(4, 4)$. That score surface intersects with the surface for the score of 1 along the line $\theta_1 = -1$. The surface for a score of 1 is near 1 when the $\boldsymbol{\theta}$-vector is $(4, -4)$ and it is near zero for the vector $(-4, 4)$. It intersects with the surface for a score of 2 along the line $\theta_2 = 1$. Thus, for the region with θ_1 greater than -1 and θ_2 greater than 1, the score of 2 is the most likely score.

The expected score surface for this test item is obtained by multiplying the score category by the probability and summing over score categories. The expected score surface for the score response surfaces shown in Fig. 4.13 is presented in Fig. 4.14.

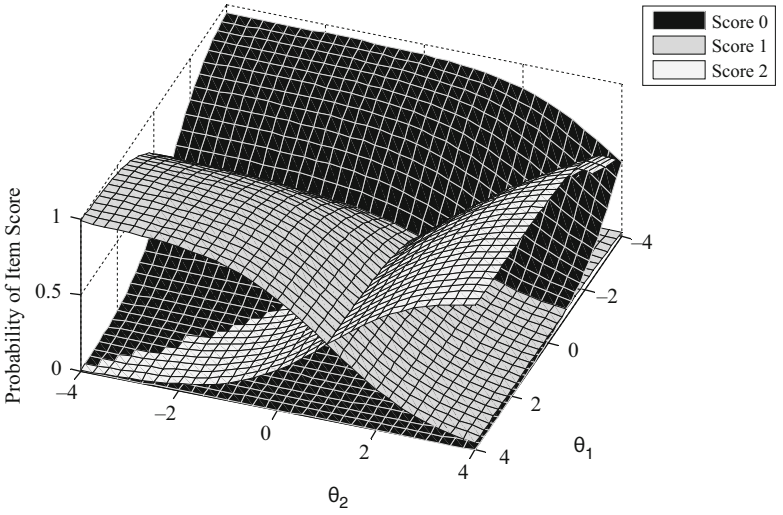

Fig. 4.13 Item response surfaces for a Kelderman and Rijkes model for a test item that requires capabilities on two dimensions to obtain a correct response – score categories of 0, 1, 2 – difficulty parameters -1 for dimension 1 and $+1$ for dimension 2

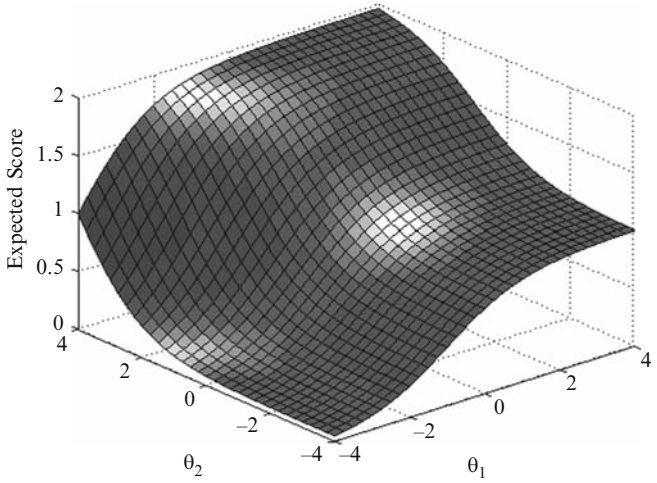

Fig. 4.14 Expected score surface for the Kelderman and Rijkes model for the test item in Fig. 4.13

This surface has an expected score near 2 when levels on both θ-coordinates are 4. The expected score drops as either θ coordinate is reduced, but there is a leveling of the surface in the region around $(4, -4)$. Examinees in this region have mastered the first task in the test item, but have very low proficiency on the second task. Therefore, their expected score is 1.

This is a very interesting model because it acknowledges that the skills and knowledge needed to achieve the highest score on a test item may be different than those needed to achieve lower scores. The user of the model must confront the challenge of identifying the particular skills and knowledge needed for each score category. This might not be too great a problem for one item, but those skills and knowledge categories need to generalize across the set of test items on a test. At this time there is not much in the research literature on the effect of misidentifying the required skills and knowledge for a test item.

4.1.2.3 Multidimensional Graded Response Model

Another approach to the multidimensional modeling of the responses to test items with more than two score categories was presented by Muraki and Carlson (1993). This model is a generalization of the unidimensional graded response model and it uses response functions that have the normal ogive form. As with the unidimensional version of this model, the multidimensional model assumes that successful accomplishment of the task specified by the test item requires a number of steps and reaching step k requires success on step $k - 1$. This type of model is also appropriate for rating scales where a rating category subsumes all previous categories. An example is rating scale for the amount of time spent on a project. If a rating category

indicating one hour was spent on a project is selected, which means that all rating categories specifying less than one hour also apply.

The parameterization of the model given here considers the lowest score on Item i to be 0 and the highest score to be m_i. The probability of accomplishing k or more steps is assumed to increase monotonically with an increase in any of the hypothetical constructs underlying the test as represented by the elements of the $\boldsymbol{\theta}$-vector. This is equivalent to dichotomizing the scale at k and scoring k or higher as a 1 and below k as 0 and fitting a dichotomous model to the result. The probability of accomplishing k or more steps is modeled by a two-parameter normal ogive model with the person parameter defined as a linear combination of the elements in the $\boldsymbol{\theta}$-vector weighted by discrimination parameters. The probability of receiving a specific score, k, is the difference between the probability of successfully performing the work for k or more steps and successfully performing the work for $k + 1$ or more steps. If the probability of obtaining an item score of k or higher at a particular $\boldsymbol{\theta}$-level is $\mathrm{P}^*(u_{ij} = k \mid \boldsymbol{\theta}_\mathrm{j})$, then the probability that an examinee will receive a score of k is

$$P(u_{ij} = k \mid \boldsymbol{\theta}_j) = P^*(u_{ij} = k \mid \boldsymbol{\theta}_j) - P^*(u_{ij} = k + 1 \mid \boldsymbol{\theta}_j), \tag{4.27}$$

where $P^*(u_{ij} = 0 \mid \boldsymbol{\theta}_j) = 1$ because doing the work for step 0 or more is a certainty for all examinees and $P^*(u_{ij} = m_i + 1 \mid \boldsymbol{\theta}_j) = 0$ because it is impossible do work representing more than category m_i. The latter probability is defined so that the probability of each score can be determined from (4.27). Samejima (1969) labels the terms on the right side of the expression as the cumulative category response functions and those on the left side of the expression as the category response function.

The normal ogive form of the graded response model is given by

$$P(u_{ij} = k \mid \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{a_i'\theta_j} + d_{i,k+1}}^{\mathbf{a_i'\theta}_j + d_{ik}} \mathrm{e}^{\frac{-t^2}{2}} \mathrm{d}t, \tag{4.28}$$

where k is the score on the item, 0, 1, …, m_i, $\mathbf{a_i}$ is a vector of item discrimination parameters, and d_{ik} is a parameter related to ease with which a person will reach the kth step of the item.

Note that the d_{ik} parameter has high positive values when it is relatively easy to obtain a particular score and large negative values when it is difficult to obtain a particular score. The d_{ik} parameters have an inverse relationship with the scores for the item. For score category 0, $d_{i0} = \infty$ and when the score category is $m_i + 1$, a value that is a point higher than actually exists on the score scale for the test item, $d_{i,m_i+1} = -\infty$. Only the values of d_{ik} from $k = 1$ to m_i are estimated in practice.

The probability of response category k can also computed from the difference of two integral expressions. This representation of the model is given in (4.29).

$$P(u_{ij} = k \mid \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_{ik}} \mathrm{e}^{\frac{-t^2}{2}} \mathrm{d}t - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_{i,k+1}} \mathrm{e}^{\frac{-t^2}{2}} \mathrm{d}t. \qquad (4.29)$$

This representation of the model makes it clear that it is based on dichotomizing the score scale for the item at different score values and using the normal ogive model to describe the probability of obtaining a score at or above that value. The probability of a particular score value, k, is the difference between the probability of being k or higher and $k + 1$ or higher.

Plots of the category response functions for a test item with four response categories, 0, 1, 2, 3, are shown in Fig. 4.15. This test item has item parameters $a_{i1} = 1.2$, $a_{i2} = .7$, $d_{i1} = .5$, $d_{i2} = -1.5$, and $d_{i3} = -2.5$. Careful study of the figure reveals that as the θs increase, the probability of the score of 0 decreases and the score of 3 increases. The intermediate scores of 1 and 2 increase, then decrease, as the θs increase.

The expected score on the item is computed by multiplying the score by the probability of the score. The expected score surface for the item shown in Fig. 4.15 is presented in Fig. 4.16. The expected score is near 0 when the elements of the $\boldsymbol{\theta}$-vector are both near -4 and it increases to near 3 when the elements of the $\boldsymbol{\theta}$-vector are both near 4. Because the multidimensional graded response model is based on a compensatory multidimensional model, each of the response probability surfaces has parallel equal probability contours. Therefore, the expected score surface also has parallel equal score contours.

The slope of this surface is dependent on both the magnitude of the elements of the $\mathbf{a}$-parameter vector and the amount of variation on the d-parameters. The

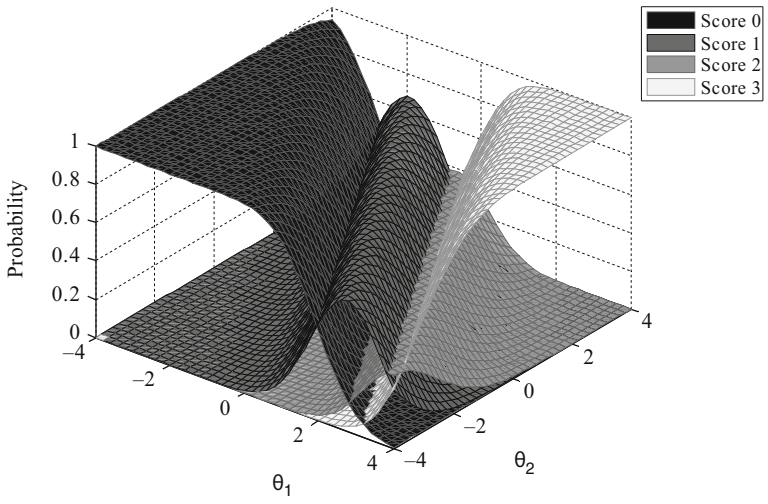

Fig. 4.15 Category response surfaces for an item with four score categories modeled by the graded response model

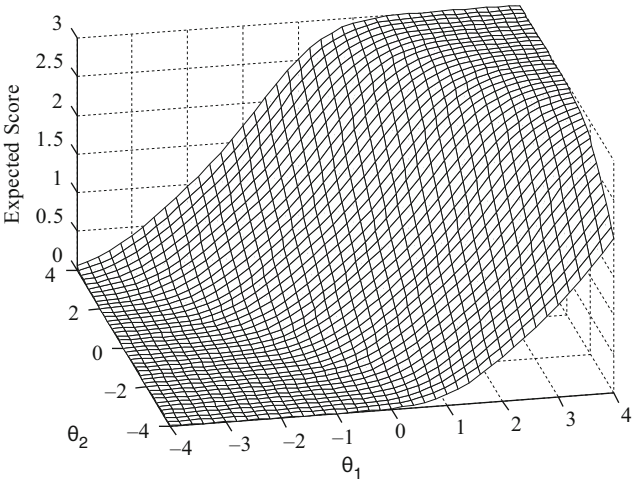

Fig. 4.16 Expected score surface for the item shown in Fig. 4.15

expected response surface is steeper when the variation in d-parameters is reduced. If the d-parameters are all equal, the item functions as a dichotomous item with scores of 0 and m.

The multidimensional graded response model has category response functions that appear similar to those from the multidimensional generalized partial credit model. A comparison of Figs. 4.11 and 4.15 will show the apparent similarity. Although, the models appear similar, they have different underlying scoring processes and work by van der Ark (2001) on the unidimensional versions of the models show that it is possible to distinguish between data sets generated from the two models.

4.2 Future Directions for Model Development

Although many different kinds of MIRT models have been proposed, it is likely that more models will be proposed as the interactions between persons and items become better understood. For example, it is likely that persons scoring performance assessments such as writing samples consider different combinations of skills and knowledge when they assign scores representing different levels of performance. It may be that the difference between 0 and 1 on the rubric for scoring a writing sample may focus on simple writing mechanics, while differences between the top score categories likely reflect differences in organization and style of writing. Such shifts in the focus of the scoring rubrics suggest that there should be vectors of **a**-parameters for each score boundary rather than a single **a**-parameter vector for the item. A merger of the Kelderman and Rijkes (1994) model and the multidimensional

versions of the generalized partial credit or graded response models may be needed to accurately represent scores from such rubrics.

It is also hypothesized that for some dichotomously scored items different examinees have different solution strategies. A student who has recently studied the topic that is the focus of the test item may be generating a response based on memory of recent assignments. Another student may need to use problem solving skills to determine how to approach the item. If situations can be identified where such hypotheses can be supported, it may be necessary to merge latent class models and MIRT models. The latent class would define the solution strategy and the vectors of θ-elements would specify the skills and knowledge needed by each strategy. Model development is continuing and it is likely that new models or variations on the existing models will appear in the research literature.

4.3 Exercises

1. Select a test item that you believe would be well modeled by a compensatory model and another test item that would be will modeled by a partially compensatory model. Describe why you believe each of the two items would be best represented by the different types of models.

2. A test item has three score categories and it is well fit by the multidimensional graded response model. The item has been calibrated assuming a three-dimensional $\boldsymbol{\theta}$-space. The item parameter estimates for the item are given in the table below. Compute the probability of each score category for persons with the following two $\boldsymbol{\theta}$-vectors – [−.5 .2 1] and [1.5-.7 0].

a_1	a_2	a_3	d_1	d_2
.5	1	1.2	.8	−1.2

3. A dichotomously scored test item is well modeled by the compensatory logistic model using three coordinate dimensions. The $\mathbf{a}$-parameter vector for the item is [.8 1.2 .3] and the d parameter is .5. What is the slope of the item response surface parallel to each coordinate axis at the point [0 0 0] in the $\boldsymbol{\theta}$-space?

4. A dichotomously scored test item is well modeled by the compensatory logistic model using two coordinate dimensions. The $\mathbf{a}$-parameter vector for the item is [1.5 .5] and the d parameter is −1. Draw the lines in the $\boldsymbol{\theta}$-plane where the probability of correct response for the item is .2 and .7. Determine the perpendicular distance between the two lines. Do the same for an item with the same d parameter, but with $\mathbf{a}$-parameter vector [1.2 1.2]. For which item are the two lines closer together? Explain why there is a difference in the distance for the two items.

5. A short quiz with five test items has been modeled using a MIRT model and the probability of correct response for each item is predicted for a student in the class. The probabilities for the five items are .9, .75, .60, .55, .5. After the quiz was administered the following scores were recorded for that student – 1 1 0 1 0. The item score vector and the probability vector have the items in the same order. What is the probability of the set of item scores for this student based on the MIRT model? What assumption of the model was important for computing the probability of the set of item scores?

6. Using the parameters for the two test items given in 4 above, compute the distance from the origin of the $\boldsymbol{\theta}$-space to the .5 equiprobable contour for each item.

7. Does the data presented in Table 4.1 support the use of a compensatory or partially compensatory MIRT model? Give the reasons for your conclusion using specific information from the table.

8. For a compensatory MIRT model with a lower asymptote parameter, c, that is not zero, what probability of correct response corresponds to $\boldsymbol{\theta}$-vectors that meet the condition $\mathbf{a}'\boldsymbol{\theta} + d = 0$?

9. A test item is well modeled by the Rasch version of the partially compensatory model. The number of elements in the $\boldsymbol{\theta}$-vector is four and all of the b-parameters for the item are equal. The observed probability of correct response for the item is .41 for a group of persons who all have $\boldsymbol{\theta}$-vectors equal to the 0 vector. What value of the b-parameters is consistent with this information?

Chapter 5
Statistical Descriptions of Item and Test Functioning

The MIRT models in Chap. 4 provide mathematical descriptions of the interactions of persons and test items. Although the parameters of these models summarize the characteristics of the items, the vectors of item parameters sometimes lack intuitive meaning. This chapter provides other statistical ways of describing the functioning of test items that may more clearly indicate the value of the test items for determining the location of individuals in the multidimensional $\boldsymbol{\theta}$-space. The ways of describing test item characteristics given here are direct extensions of the descriptive information for UIRT models described in Chap. 2.

5.1 Item Difficulty and Discrimination

The UIRT measures of item difficulty and discrimination are directly related to the characteristics of the item characteristic curve (ICC). The difficulty parameter indicates the value of θ that corresponds to the point of steepest slope for the ICC. The discrimination parameter is related to the slope of the ICC where it is steepest. These two descriptive statistics for test items can be generalized to the MIRT case, but there are some complexities to the process. The slope of a surface is dependent on the direction of movement along the surface so the point of steepest slope depends on the direction that is being considered. For example, the contour plot of the item response surface for a test item that is well modeled by the multidimensional extension of the two-parameter logistic (M2pl) model is shown in Fig. 5.1. The solid arrow in the figure shows a direction that is parallel to the equi-probable contours for the surface, the set of points in the $\boldsymbol{\theta}$-space that yield the same probability of correct response to the test item. Over the length of the arrow, there is no change in probability so the slope in that direction is zero. The dashed arrow is in a direction with substantial change in probability – .4 over the length of the arrow. Because the length of the arrow is about one unit, the slope in that direction is about .4.

At each point in the $\boldsymbol{\theta}$-space, there is a direction that has the maximum slope from that point. If the entire $\boldsymbol{\theta}$-space is considered and the slopes in all directions at each point are evaluated, there is a maximum slope overall for the test item. The value of the maximum slope would be a useful summary of the capabilities of the

M.D. Reckase, *Multidimensional Item Response Theory*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-89976-3_5,

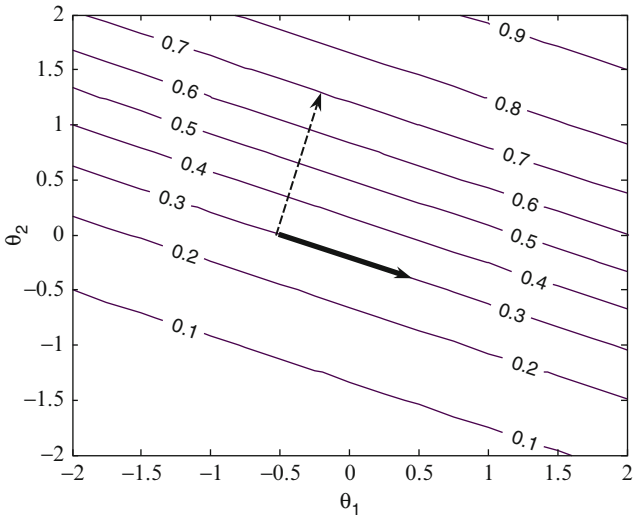

Fig. 5.1 *Arrows* showing differences in slope for an M2pl item with parameters $a_1 = .5, a_2 = 1.2$, $d = -.6$

test item for distinguishing between $\boldsymbol{\theta}$-points in the direction of greatest slope. It would also be helpful to know the relationship between the **a**-parameter vector and the values of the slopes at a point in the $\boldsymbol{\theta}$-space. Knowing the relationship will allow the evaluation of the usefulness of the test item for differentiating between $\boldsymbol{\theta}$-points at different locations in the space using estimates of the item parameters.

The b-parameter in UIRT is a measure of the distance from the 0-point of the θ-scale to the value on the scale that is below the point of steepest slope for the ICC. The sign of the b-parameter indicates the direction from the 0-point of the θ-scale to the location of the point of steepest slope. By convention, negative signs indicate distances to the left of the 0-point and positive signs indicate distances to the right of the 0-point (see Sect. 2.1.1 for a discussion of the difficulty in UIRT models). It would be useful to have a similar indicator for test items that are described using multidimensional models. The parallel measure of item difficulty would be the distance from the origin of the $\boldsymbol{\theta}$-space (i.e., the $\mathbf{0}$-vector) to the $\boldsymbol{\theta}$-point that is below the point of steepest slope for the surface. The sign associated with this distance would indicate the relative position of the $\boldsymbol{\theta}$-point to the origin of the $\boldsymbol{\theta}$-space. Also, it would be useful if the distance to this point were related to the d-parameter in the model.

The statistics that correspond to the a and b-parameters for the UIRT models are derived here for the M2pl model. A similar derivation can be used for other compensatory models, but they do not generalize to the partially compensatory models. The basic concepts apply to the partially compensatory models, but the corresponding statistical summaries do not result in simple mathematical expressions.

The goal of developing these statistics is to determine the point of steepest slope for the surface and the distance from the origin of the $\boldsymbol{\theta}$-space to that point. Because

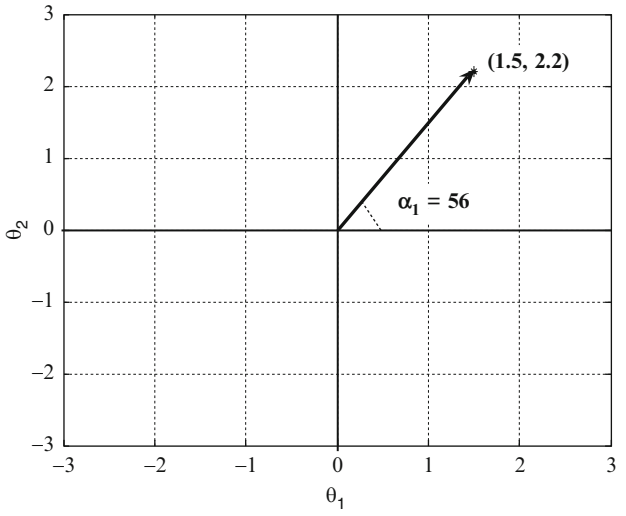

Fig. 5.2 Polar coordinate representation of a point in a two-dimensional $\boldsymbol{\theta}$-space – the length of *arrow* is 2.66 θ-coordinate units

this conceptualization uses the distance from the origin of the $\boldsymbol{\theta}$-space as a reference point for the measure of item difficulty, it is helpful to reparameterize the M2pl model using a polar coordinate representation. That is, instead of representing each point in the $\boldsymbol{\theta}$-space by a vector of θ-coordinates, each point is represented by a vector of angles from each coordinate axis and a distance from the origin. A two-dimensional representation of this reparameterization is given in Fig. 5.2. The same conceptual framework can be used to generalize the reparameterization to an n-dimensional space.

In Fig. 5.2, the location of a point in the $\boldsymbol{\theta}$-space is indicated by the two coordinates, $\theta_1 = 1.5$ and $\theta_2 = 2.2$. That same point can also be represented by a distance (i.e., the bold arrow) of 2.66 θ-units and a direction of 56° from the θ_1-axis. The distance is computed using the standard distance formula and the angle is determined from the right triangle trigonometric formula for the cosine of an angle. In this two-dimensional case, the angle between the arrow and the θ_2-axis can be determined by subtraction as $90 - 56 = 34°$. In m dimensions, $m - 1$ angles can be computed from the θ-coordinates using trigonometric relationships. The mth angle is mathematically determined because the sum of squared cosines must equal 1. Given the distance to the point and the directions from the axes, the values of the coordinates of the point on each of the axes can be recovered using the trigonometric relationship

$$\theta_v = \zeta \cos \alpha_{v,} \tag{5.1}$$

where θ_v is the coordinate of the point on dimension v, ζ is the distance from the origin to the point, and α_v is the angle between the vth axis and the line from the origin to the point.

The expression on the right side of (5.1) can be substituted for the θ-coordinates in the exponent of the M2pl model. After the substitution, the model is given by

$$P\left(U_{ij}=1|\zeta_j,\mathbf{a_i},\boldsymbol{\alpha}_\mathbf{j},d_i\right)=\frac{\mathrm{e}^{\left(\zeta_j\sum\limits_{\ell=1}^{m}a_{i\ell}\cos\alpha_{j\ell}\right)+d_i}}{1+\mathrm{e}^{\left(\zeta_j\sum\limits_{\ell=1}^{m}a_{i\ell}\cos\alpha_{j\ell}\right)+d_i}}, \tag{5.2}$$

where ζ_j is a scalar parameter for Person j that indicates the distance from the origin to the location of the person, $\boldsymbol{\alpha}_\mathbf{j}$ is the vector of angles between the coordinate axes and the line from the origin to the point representing the location of Person j in the solution space, and the other symbols have the same meaning as for the M2pl model. The $\boldsymbol{\alpha}_\mathbf{j}$ vector has m elements, but only $m-1$ of them need to be estimated because the squared cosines must sum to 1.

The form of the model given in (5.2) is useful because the slope in a direction specified by the $\boldsymbol{\alpha}$-vector can be determined by taking the partial derivative of the model equation with respect to the single scalar variable, ζ_j. This partial derivative is given in (5.3).

$$\frac{\partial P\left(U_{ij}=1|\zeta_j,\mathbf{a_i},\boldsymbol{\alpha}_\mathbf{j},d_i\right)}{\partial\zeta_j}=P_{ij}Q_{ij}\sum_{\ell=1}^{m}a_{i\ell}\cos\alpha_{j\ell}. \tag{5.3}$$

In (5.3), the P_{ij} and Q_{ij} symbols are abbreviated notation for the probability of correct and incorrect response, respectively, for Person j on Item i. This result shows that the slope at a location in the solution space is dependent on the probability of correct response at that location, the elements of the $\mathbf{a}$-parameter vector, and the angles with the axes indicated by the $\boldsymbol{\alpha}$-vector. If the angle with an axis is 0°, the corresponding cosine is 1 and all of the other cosines are 0. The slope along an axis simplifies to $P_{ij}Q_{ij}a_{i\ell}$, for coordinate axis ℓ.

To determine the steepest slope in the direction specified by the $\boldsymbol{\alpha}$-vector, the second derivative of the item response function is taken with respect to ζ_j and the result is set equal to zero and solved for the value of ζ_j. The second derivative is given in (5.4).

$$\frac{\partial^2 P\left(U_{ij}=1|\zeta_j,\mathbf{a_i},\boldsymbol{\alpha}_\mathbf{j},d_i\right)}{\partial\zeta_j^2}=\left(\sum_{\ell=1}^{m}a_{i\ell}\cos\alpha_{j\ell}\right)^2P_{ij}\left(1-3P_{ij}+2P_{ij}^2\right). \tag{5.4}$$

There are three solutions when (5.4) is set equal to 0, but only one of them results in a finite value of ζ_j. That solution is when $P_{ij}=.5$. The probability is .5 when the exponent of (5.2) is 0. Solving for the value of ζ_j that results in 0 gives the location along the line in the direction specified by the $\boldsymbol{\alpha}$-vector where the surface has maximum slope. The result is

$$\frac{-d_i}{\sum_{\ell=1}^{m} a_{i\ell} \cos \alpha_{j\ell}}, \tag{5.5}$$

where all of the symbols have been previously defined. The location of the point of maximum slope along a particular axis is simply $-d/a_{i\ell}$ because all of the cosines will be 0 except for the axis ℓ being considered. For that axis, the cosine is 1.

Substituting the expression in (5.5) for ζ_j in (5.2) results in a probability of a correct response of .5 for a person located along the line from the origin at the point that gives the steepest slope. As a result, the value of the slope at the point of steepest slope in the direction specified by the $\boldsymbol{\alpha}$-vector is

$$\frac{1}{4} \sum_{\ell=1}^{m} a_{i\ell} \cos \alpha_{j\ell}. \tag{5.6}$$

To determine the direction of steepest slope from the origin of the $\boldsymbol{\theta}$-space, the expression in (5.6) is differentiated with respect to cos $\boldsymbol{\alpha}$ and solved for 0. This is done under the constraint that the sum of the squared cosines is equal to 1. The result is the system of equations given in (5.7).

$$a_{i\ell} - a_{im} \frac{\cos \alpha_{i\ell}}{\cos \alpha_{im}} = 0, \quad \text{for } \ell = 1, 2, \ldots, m-1, \tag{5.7}$$

where $\cos^2 \alpha_{im} = 1 - \sum_{k=1}^{m-1} \cos^2 \alpha_{ik}$. The solution for the system of equations is given by

$$\cos \alpha_{i\ell} = \frac{a_{i\ell}}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}. \tag{5.8}$$

The corresponding angles are given by taking the arccosine of the cosine of α. These angles and cosines are characteristics of the item. They indicate the direction from the origin of the $\boldsymbol{\theta}$-space to the point in the $\boldsymbol{\theta}$-space that has the greatest slope considering all possible directions. The cosines specified by (5.8) are sometimes called *direction cosines*.

The distance from the origin to the point of steepest slope in the direction specified by (5.8) can be obtained by substituting the results of (5.8) for the cos α in (5.5). The result is

$$B_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}. \tag{5.9}$$

The symbol B_i is used here to represent the multidimensional difficulty of the test item. Sometimes this item characteristic is represented by MDIFF, but B is used here to more clearly make the connection to the unidimensional b-parameter because it

has an equivalent interpretation to that of the b-parameter in UIRT models. That is, high positive values of B indicate difficult items (i.e., those that require high values of the elements of $\boldsymbol{\theta}$ to yield a probability of a correct response greater than .5). Low values of B indicate items with a high probability of correct response for the levels of $\boldsymbol{\theta}$ that are usually observed.

This interpretation of B applies only to the direction specified by the $\boldsymbol{\alpha}$-vector. Thus, this analysis of the characteristics of a test item results in two descriptive measures. One is an indication of the difficulty of the test item (i.e., B) and the other is a description of the combination of the coordinate axes that is most differentiated by the test item (i.e., $\boldsymbol{\alpha}$). This combination is indicated by the direction of steepest slope from the origin of the $\boldsymbol{\theta}$-space.

A value that is analogous to the discrimination parameter from the UIRT model can also be defined. In UIRT, the discrimination parameter is related to the slope at the point of steepest slope for the ICC. The equivalent conceptualization for the discrimination parameter in the MIRT case is the slope of the item response surface at the point of steepest slope in the direction from the origin of the θ-space. This slope can be determined by substituting (5.8) into (5.6). The slope is $1/4$ the value presented in (5.10). As with the unidimensional IRT models, the constant $1/4$ is not included in the expression resulting in a multidimensional discrimination index of

$$A_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2}. \tag{5.10}$$

A_i is the multidimensional discrimination for Item i. In some articles, the term MDISC_i is used instead of A_i. Here, A_i is used to emphasize the connection to the a-parameter in the unidimensional models. Note that A_i has the same mathematical form as the term in the denominator of (5.9). Therefore, another expression for the multidimensional difficulty is $B_i = -d_i/A_i$.

Two examples are provided to give intuitive meaning to these descriptive indices of multidimensional items. Figure 5.3 provides the equi-probable contours for two test items that are described in a two-dimensional coordinate system. The parameters for the two items are given in Table 5.1 along with the multidimensional discrimination and difficulty, and the angles with the coordinate axes for the direction of maximum slope. Several features can be noted from the plots. First, the contour lines are closer together for Item 1 than for Item 2. This shows that Item 1 is more discriminating than Item 2 because the probabilities are changing more quickly with change in location of $\boldsymbol{\theta}$-points. This is also shown by the value of A_i for the two items. Item 1 has a larger value for the multidimensional discrimination.

A second feature of the contour plots is that the contour lines have different orientations to the coordinate axes. Each plot also contains an arrow that shows the direction of steepest slope from the origin of the $\boldsymbol{\theta}$-space. The angles with the coordinate axes for the arrows are given by the α-values. For Item 1, the arrow is about 23° from the θ_1-axis and about 67° from the θ_2-axis. The angles are quite different for the arrow for Item 2. Note that the arrows are also pointing in different directions.

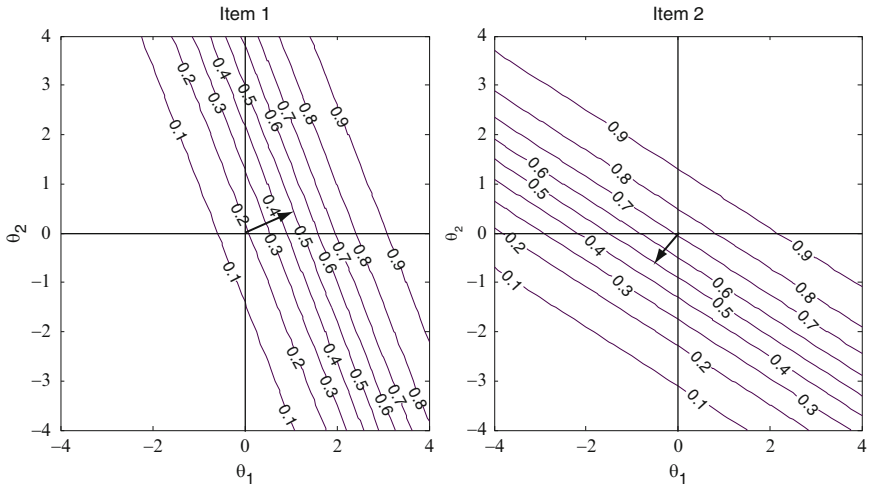

Fig. 5.3 Representation of the characteristics of the items with parameters in Table 5.1

Table 5.1 Item parameters and multidimensional statistics for two test items

Item	a_1	a_2	d	A	B	α_1	α_2
1	1.2	.5	−1.5	1.30	1.15	22.62	67.38
2	.6	1	.9	1.17	−.77	59.04	30.96

The arrows stretch from the origin of the $\boldsymbol{\theta}$-space to the line of the .5-probability contour. When the arrow is in the lower left quadrant of the space, the item is fairly easy. When the arrow is in the upper right quadrant, the item is fairly hard. The only way that the arrow will point in the direction of one of the other quadrants is if one of the a-parameters is negative. The length and direction of the arrow is given by B_i. Negative values generally indicate easy items and positive values hard items.

If the item response surfaces for multiple items are plotted as equi-probable contour plots on the same set of axes, the characteristics of the individual items will be difficult to discern because of the number of intersecting equi-probable lines. One way to show the same information in a less cluttered way is to represent each item by an arrow with the base of the arrow at the point of maximum slope along a line from the origin of the $\boldsymbol{\theta}$-space. The arrow points up slope in the direction along the line from the origin. The length of the arrow can be used to represent the discriminating power, A_i, for the item. The distance from the origin to the base of the arrow indicates the difficulty, B_i, for the item and the direction, $\boldsymbol{\alpha}_i$, of the arrow shows the direction of greatest positive change in slope for the item. This type of representation for the two items in Table 5.1 is given in Fig. 5.4. Using these conventions, a number of items can be displayed graphically when the number of coordinate axes is two or three. Figure 5.5 gives the arrow representation for 45 items in a three-dimensional space. This representation of the items makes it clear that the items form three fairly distinct sets that tend to have the steepest slope in the same direction.

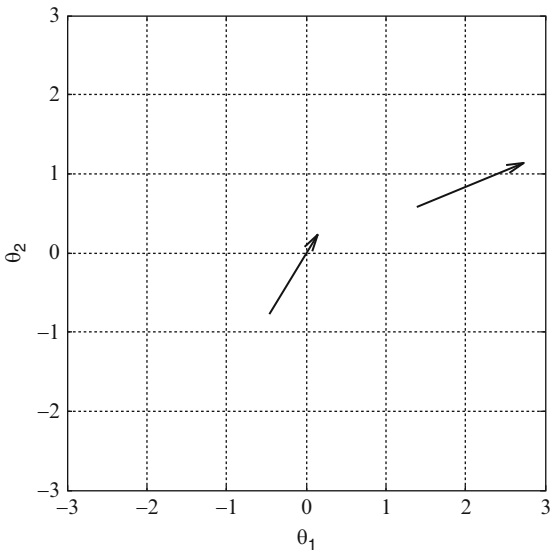

Fig. 5.4 Representation of items with parameters in Table 5.1 as *arrows*

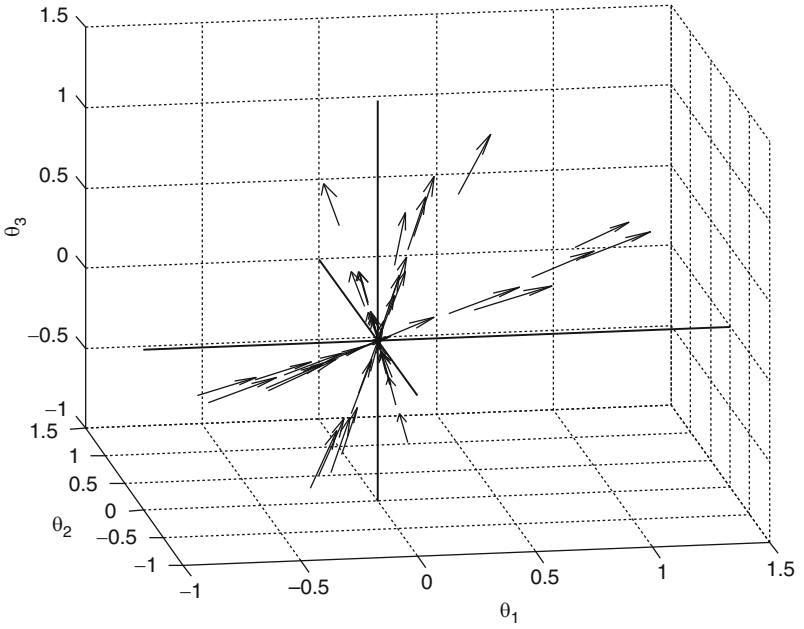

Fig. 5.5 Representation of 45 items by *arrows* in a three-dimensional space

The description of test items using the concepts of multidimensional difficulty, multidimensional discrimination, and direction of steepest slope in the multidimensional space can also be used with polytomous test items. Muraki and Carlson (1993) derive these statistics for the multidimensional graded response model. The

development results in the same expressions for multidimensional discrimination, A_i, and the vector of angles, $\boldsymbol{\alpha}$. They also developed the equivalent to (5.9) for the step difficulty for an item (see (4.28)).

$$B_{ik} = \frac{-d_{ik}}{\sqrt{\sum_{\ell=1}^{m} a_{i\ell}^2}}, \tag{5.11}$$

where B_{ik} is the step difficulty for the kth step of the graded response item and d_{ik} is the step parameter as defined in the model. The other terms have been previously defined.

The same procedures can be used to define the multidimensional statistics for the item whenever the exponential term in the model is in the form $\mathbf{a}\boldsymbol{\theta}' + d$. This is also the case when the limits of integration for a normal ogive model have the same form. Partially compensatory models do not have this form; therefore, these statistical descriptions of characteristics of the test items do not apply to test items that are described by partially compensatory models. It may be possible to derive similar statistics for test items model by the partially compensatory models, but they have not been developed at the time this book was written.

5.2 Item Information

The concept of item information that is used in UIRT can also be generalized to the multidimensional case. The definition of information in the multidimensional case is the same as that given in (2.42) for the unidimensional case – the squared slope of the regression of the item score on the $\boldsymbol{\theta}$ divided by the variance of the item score at θ. There is a complication, however. At each point in the $\boldsymbol{\theta}$-space, the slope of the multidimensional item response surface differs depending on the direction of movement from the point. The two arrows in Fig. 5.1 give examples of the differences in slope for movements in different directions. The solid arrow shows movement in a direction that has a slope of 0. The dashed arrow has a much higher positive slope. This implies that the information provided by a test item about the difference between nearby points in the $\boldsymbol{\theta}$-space depends on the orientation of the points (i.e., the direction of movement from one point to the other).

To accommodate the change in slope with direction taken from a point in the $\boldsymbol{\theta}$-space, the definition of item information is generalized to

$$I_{\alpha}(\boldsymbol{\theta}) = \frac{[\nabla_{\alpha} P(\boldsymbol{\theta})]^2}{P(\boldsymbol{\theta})Q(\boldsymbol{\theta})}, \tag{5.12}$$

where $\boldsymbol{\alpha}$ is the vector of angles with the coordinate axes that defines the direction taken from the $\boldsymbol{\theta}$-point, ∇_{α} is the directional derivative or gradient, in the direction $\boldsymbol{\alpha}$, and the other symbols as previously defined. Equation (5.12) represents the

information for one item at one location in the $\boldsymbol{\theta}$-space, so the item and person subscripts have not been included to more clearly show the general structure of the expression.

The directional derivative for the item response surface is given by

$$\nabla_{\boldsymbol{\alpha}} P(\boldsymbol{\theta}) = \frac{\partial P(\boldsymbol{\theta})}{\partial \theta_1} \cos \alpha_1 + \frac{\partial P(\boldsymbol{\theta})}{\partial \theta_2} \cos \alpha_2 + \cdots + \frac{\partial P(\boldsymbol{\theta})}{\partial \theta_m} \cos \alpha_m. \tag{5.13}$$

If the MIRT model being considered is the multidimensional extension of the two-parameter logistic model given in (4.5), the directional derivative is

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}} P(\boldsymbol{\theta}) = {} & a_1 P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \cos \alpha_1 \\ & + a_2 P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \cos \alpha_2 + \cdots + a_m P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \cos \alpha_m. \end{aligned} \tag{5.14}$$

This expression can be presented more compactly as

$$\nabla_{\boldsymbol{\alpha}} P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \sum_{v=1}^{m} a_v \cos \alpha_v. \tag{5.15}$$

Substituting (5.15) into (5.12) yields

$$I_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \frac{\left[P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \sum_{v=1}^{m} a_v \cos \alpha_v \right]^2}{P(\boldsymbol{\theta}) Q(\boldsymbol{\theta})} = P(\boldsymbol{\theta}) Q(\boldsymbol{\theta}) \left(\sum_{v=1}^{m} a_v \cos \alpha_v \right)^2. \tag{5.16}$$

When the MIRT model contains only two dimensions, the information in the direction specified by the $\boldsymbol{\alpha}$-vector can be represented by an information surface. The height of the surface above the θ-plane indicates the amount of information at each location in the plane. Information surfaces for the test item represented in the contour plot shown in Fig. 5.1 are shown in Fig. 5.6. The surfaces show the information in three directions – angles of 0, 67.38, and 90° with the θ_1-axis.

The angle of 67.38° is the direction from the θ_1-axis when the item response surface has the steepest slope along a line from the origin of the space. In all three cases, the information surfaces have the shape of a ridge that has its highest region over the .5 equiprobable contour line for the test item. That is the line in the $\boldsymbol{\theta}$-space where the slope of the item response surface is the greatest as well. Of all of the possible directions in the space, the direction of steepest slope has the highest ridge for the information function. The low maximum height for the ridge in the panel on the left of Fig. 5.6 shows that the test item is relatively ineffective at distinguishing nearby points when the direction between the points is parallel to the θ_1-axis. However, the test item is quite effective at distinguishing points on either side of the .5 equiprobable contour in the direction of steepest slope for the item response surface – 67.38° to the θ_1-axis. This is shown by the high ridge in the middle panel

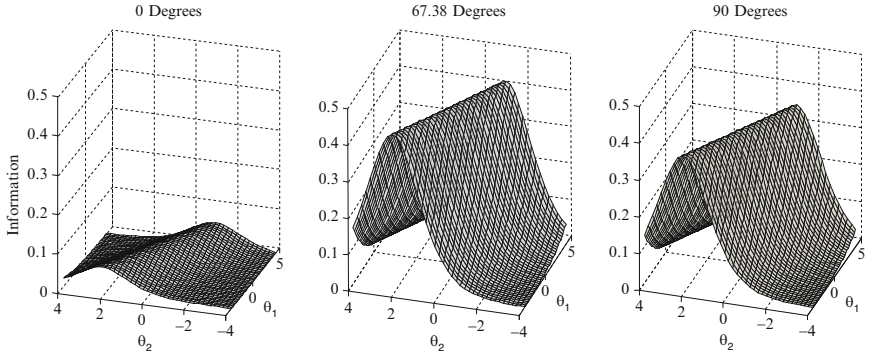

Fig. 5.6 Information surfaces for a M2PL test item with $a_1 = .5$, $a_2 = 1.2$, and $d = -.6$ in three directions

of Fig. 5.6. The capability of the item for distinguishing in a direction parallel to the θ_2-axis is between the other two directions as shown by the right panel in the figure. The maximum height of the surface is between the maximum heights of the surfaces for the other two directions.

If the direction of steepest slope from (5.8) is substituted for the $\cos \alpha_k$ in (5.16), the result is the information for the test item in the direction of maximum slope. That expression is given by

$$I_{\alpha \max}(\boldsymbol{\theta}) = P(\boldsymbol{\theta})Q(\boldsymbol{\theta}) \sum_{k=1}^{m} a_k^2 = P(\boldsymbol{\theta})Q(\boldsymbol{\theta})A^2, \tag{5.17}$$

where A is the multidimensional discrimination for the item. In this case, the information function has exactly the same form as that for the two-parameter logistic model described in Chap. 2. For the example in Fig. 5.6, $A = 1.3$ and because the term $P(\boldsymbol{\theta})Q(\boldsymbol{\theta})$ has its maximum when $P(\boldsymbol{\theta}) = .5$, the maximum information is $.5 \times .5 \times 1.3^2 = .425$.

Because the information surface is different in every direction that is selected, it is difficult to get a sense of the overall information provided by an item. One approach to addressing this issue was given by Reckase and McKinley (1991). For a grid of points selected in the $\boldsymbol{\theta}$-space, the information was determined in directions from the θ_1-axis in 10° increments. The results were plotted as lines radiating from the $\boldsymbol{\theta}$-points in the selected directions with the length of the line indicating the amount of information. These plots have sometimes been labeled "clam shell" plots because the sets of lines often look like the back of a clam shell. Figure 5.7 shows a clam shell plot for the information provided by the test item shown in Fig. 5.6.

Careful examination of Fig. 5.7 will show that the lines are longest for $\boldsymbol{\theta}$-points that fall near the .5 equiprobable contour line and that the longest line in each set of lines is at 70°, the angle closest to the direction of maximum slope from the origin of the $\boldsymbol{\theta}$-space. When the $\boldsymbol{\theta}$-points are far from the .5 contour line, the information is very low and the amount is represented by a point on the graph. Clam shell

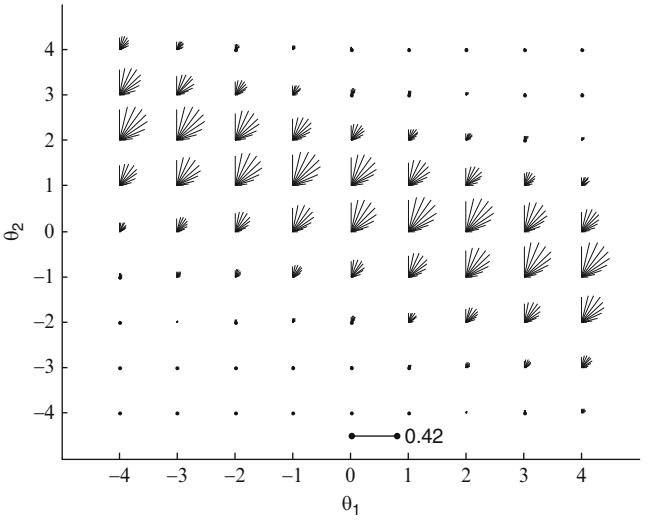

Fig. 5.7 Information for a M2PL test item with $a_1 = .5$, $a_2 = 1.2$, and $d = -.6$ at equally spaced points in the $\boldsymbol{\theta}$-space for angles from 0° to 90° at 10° intervals

plots can be used to represent information in many directions from each point in the $\boldsymbol{\theta}$-space for any of the MIRT models. Unfortunately, the plots can only be produced for two-dimensional solutions. Three-dimensional plots are possible, but the information pattern is difficult to see from such plots because lines from one point cover those from another point. There is not yet a satisfactory solution for graphically representing the information from items when the number of coordinate dimensions is greater than two. Although not solving the problem of representing test characteristics in high dimensions, Ackerman (1996) and Ackerman et al. (2003) provide other ways to graphically represent the characteristics of tests and test items when they are modeled using two or three-dimensional MIRT models.

5.3 MIRT Descriptions of Test Functioning

The UIRT procedure used to represent the functioning of sets of items that are scored together to represent the performance of a person, usually called tests, can be used when the tests are analyzed using MIRT models. Most of these procedures can be represented graphically when the item response data are modeled with two θ-coordinates. The procedures are the same for higher numbers of coordinate dimensions, but graphic representations are not available.

A common UIRT approach for showing the characteristics of a test is the *test characteristic curve* (see Sect. 2.2.1). This curve is the regression of the sum of the item scores on θ. This regression function can easily be generalized to the multidimensional case. The *test characteristic surface* (TCS) is the regression of the sum

of the item scores on the $\boldsymbol{\theta}$-vector. The mathematical expression for the test characteristic surface, or TCS, is exactly the same as for UIRT models (see (2.38)) except that the expectation is conditioned on the $\boldsymbol{\theta}$-vector instead of the unidimensional value of θ. The MIRT expression for the TCS for a test composed of dichotomously scored test items is given by

$$E\left(y_j|\boldsymbol{\theta}_\mathbf{j}\right) = E\left(\sum_{i=1}^{n} u_{ij}|\boldsymbol{\theta}_\mathbf{j}\right) = \sum_{i=1}^{n} E\left(u_{ij}|\boldsymbol{\theta}_\mathbf{j}\right) = \sum_{i=1}^{n} P\left(u_{ij}|\boldsymbol{\theta}_\mathbf{j}\right). \tag{5.18}$$

The TCS is simply the sum of the item characteristic surfaces for the items in the test. For tests composed of polytomously scored items, or mixtures of dichotomously and polytomously scored items, all of the terms in (5.18) still hold except the one at the far right. The TCS is the sum of the expected scores on the test items included in the test conditional on the $\boldsymbol{\theta}$-vectors.

An example of the TCS for a test is given using the item parameter estimates provided in Table 5.2. These parameter estimates were obtained from a two-dimensional solution using the program NOHARM. Detailed information about this program is given in Chap. 6. The TCS for this set of items is shown using two different graphic representations, (a) and (b), in Fig. 5.8. For points in the $\boldsymbol{\theta}$-space near $(-4, -4)$, the expected number-correct score on the set of test items is near 0. With increase in either θ-coordinate, the expected number-correct score increases until it approaches the maximum of 20 near $(4, 4)$. The surface appears similar to an item characteristic surface (see panel a), but the equal score contours are not straight lines (see panel b) as they are for an item characteristic surface (Fig. 4.5).

Another way of summarizing the characteristics of the set of test items in a test is to indicate the orientation of the unidimensional θ-scale in the multidimensional

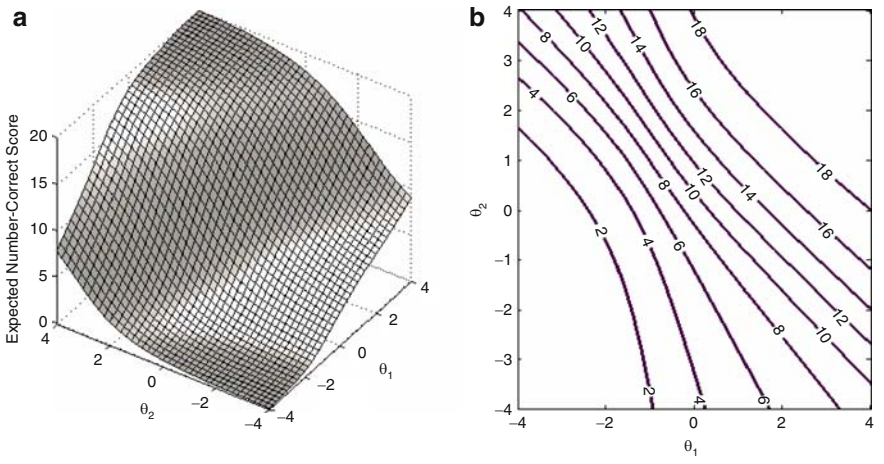

Fig. 5.8 Test characteristic surface for the test items in Table 5.2. Plot (**a**) shows the surface and plot (**b**) shows the equal score contours for the surface

Table 5.2 Item parameter estimates from a two-dimensional compensatory model

Item number	a_1	a_2	d
1	.79	1.36	−.90
2	.93	1.38	−1.20
3	.58	.38	1.00
4	.87	.87	−.97
5	.83	.79	−1.08
6	.31	.99	−1.53
7	.60	.48	−.61
8	.60	.87	−.60
9	1.64	.15	1.24
10	1.11	1.30	−.69
11	.53	.97	−1.31
12	1.26	.39	.92
13	2.37	.00	2.49
14	1.17	1.76	−.06
15	.96	1.26	−.48
16	.56	.46	−.82
17	1.17	.20	1.11
18	.63	.26	.66
19	1.01	.47	−.15
20	.81	.77	−1.08

space. This is the line in the multidimensional space that represents the unidimensional scale. The projections of the $\boldsymbol{\theta}$-points in the multidimensional space gives an estimate of the unidimensional θ that would result if the response data from the test items were analyzed using a unidimensional IRT model. Wang (1985, 1986) derived that the unidimensional θ-line corresponding to the θ-estimates from a set of test items was related to the characteristics of the matrix of discrimination parameters for the compensatory MIRT model, $\mathbf{a}$. The orientation of the unidimensional line in the $\boldsymbol{\theta}$-space is given by the eigenvector of the $\mathbf{a}'\mathbf{a}$ matrix that corresponds to the largest eigenvalues of that matrix. Wang labeled the unidimensional θ that is estimated in this way as the *reference composite* for the test.

For the $\mathbf{a}$-matrix specified by the middle two columns of Table 5.2, $\mathbf{a}'\mathbf{a}$ results in the matrix $\begin{bmatrix} 21.54 & 12.82 \\ 12.82 & 15.87 \end{bmatrix}$. The diagonal values in this matrix are the sum of the squared a-elements from the columns of the $\mathbf{a}$-matrix. The off-diagonal values are the sums of the cross-products of the a-elements from different columns. The eigenvalues for this matrix are 31.84 and 5.57. Note that the sum of the eigenvalues is the same as the sum of the diagonal elements. The eigenvector that corresponds to the larger of the two eigenvalues is $\begin{bmatrix} .7797 \\ .6292 \end{bmatrix}$. The sum of the squared elements of the eigenvector is equal to 1; therefore, the elements of the eigenvector can be considered as direction cosines. These direction cosines give the orientation of the reference composite with the coordinate axes of the $\boldsymbol{\theta}$-space.

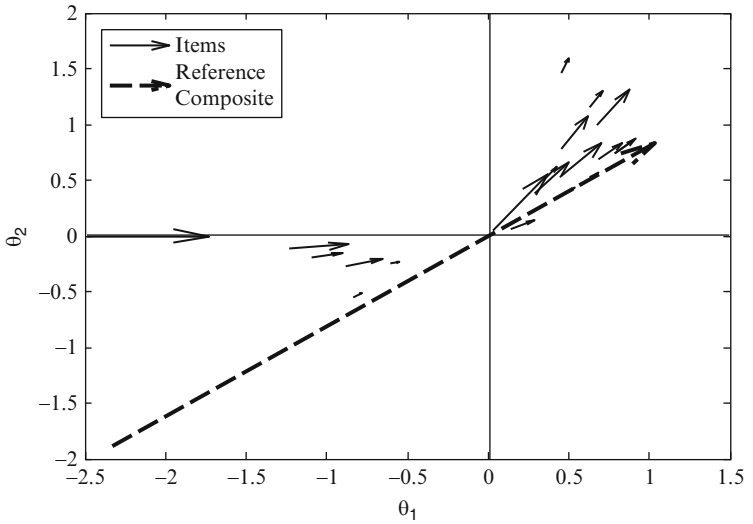

Fig. 5.9 Item *arrows* and reference composite for the item parameters in Table 5.1

Because the test items represented in the **a**-matrix were calibrated assuming a two-dimensional coordinate space, the relationship between the reference composite and the arrow representation of test items can be shown graphically. Figure 5.9 shows the arrows representing each item and the large, bold, dashed arrow representing the reference composite.

The angle between the reference composite and the coordinate axes can be determined by taking the arccosine of the elements of the eigenvector. In this example, the reference composite has an angle of approximately 39° with the θ_1 axis and 51° with the θ_2 axis. This orientation results from the slightly higher **a**-parameters related to dimension 1 than dimension 2. It is useful to compare the orientation of the reference composite with the contour plot in Fig. 5.8. The reference composite tends to be oriented along the direction of steepest slope from the origin of the $\boldsymbol{\theta}$-space for the test characteristic surface. This is not a formal relationship like that for individual test items modeled with a compensatory MIRT model because the form of the test characteristic surface is much more complex than the item characteristic surface.

Another way of showing the relationship between the unidimensional estimates of θ and the multidimensional solution is to determine the orthogonal rotation of the unidimensional θs that best matches the multidimensional $\boldsymbol{\theta}$s. The more general case of rotating one set of multidimensional $\boldsymbol{\theta}$s to match a corresponding set in another solution space will be discussed in detail in Chap. 7. The solution can be found in a number of ways. An approach with a convenient conceptual framework is used here. Suppose that a sample of two-dimensional $\boldsymbol{\theta}$-vectors used to generate item response data using the parameters given in Table 5.2 is represented by the $2{,}000 \times 2$ matrix of coordinates, $\boldsymbol{\theta}_\mathbf{t}$. These are the "true" coordinates for the examinees. From

these $\boldsymbol{\theta}$-vectors and the item parameters in Table 5.2, a 2,000 × 20 item response matrix is generated by comparing the matrix of computed probabilities of correct response for each person to each item to a matrix of uniform random numbers. If the random number is less than the computed probability, an item score of 1 is assigned. Otherwise, a 0 item score is assigned. The matrix of item scores is then used to calibrate the items and estimate θs using the unidimensional two-parameter logistic model (in this case using BILOG-MG). Those θs are used to create a $2{,}000 \times 2$ matrix of the form $\begin{bmatrix} \theta_1 & 0 \\ \theta_2 & 0 \\ \vdots & \vdots \\ \theta_{2{,}000} & 0 \end{bmatrix}$, where the index of the θs represent an examinee identification number. This matrix is represented by $\boldsymbol{\theta}_\mathbf{e}$.

To determine the orthogonal rotation that best matches the estimates in $\boldsymbol{\theta}_\mathbf{e}$ to the generating values, $\boldsymbol{\theta}_\mathbf{t}$, the singular value decomposition of the $\boldsymbol{\theta}_\mathbf{t}'\boldsymbol{\theta}_\mathbf{e}$ is determined as shown in (5.19).

$$\boldsymbol{\theta}_\mathbf{t}'\boldsymbol{\theta}_\mathbf{e} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}', \tag{5.19}$$

where $\mathbf{U}$, $\boldsymbol{\Sigma}$, and $\mathbf{V}$ are orthogonal matrices. The rotation matrix, $\mathbf{R}$, is given by $\mathbf{V}\mathbf{U}'$ and the rotated solution is $\boldsymbol{\theta}_\mathbf{e}\mathbf{R}$.

An example of the rotation of the unidimensional estimates to best match the two dimensional solution is shown in Fig. 5.10. The results in the figure show a scatter plot of the θ-coordinates for the two-dimensional $\boldsymbol{\theta}$-vectors used to generate the item response data and the corresponding line for the unidimensional θs after

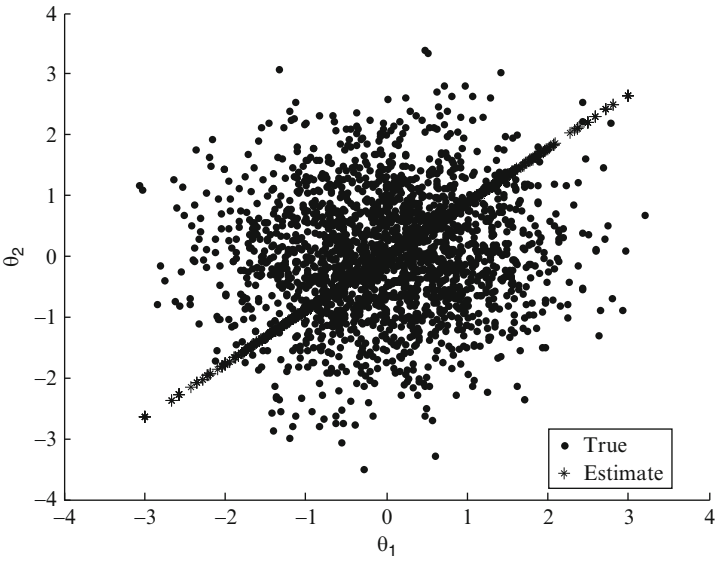

Fig. 5.10 Relationship between unidimensional estimates and two-dimensional $\boldsymbol{\theta}$-points

rotation. The rotation matrix is given by $\begin{bmatrix} .75 & .66 \\ -.66 & .75 \end{bmatrix}$. This matrix corresponds to an angle of 42° with the θ_1 axis after rotation and 48° with the θ_2 axis. These values are slightly different than the theoretical values of 39° and 51° respectively because of estimation error in the unidimensional θ estimates. Note that this rotation gives the same result as the projection of the two-dimensional points onto the unidimensional θ estimates. The rotation minimizes the sum of squared distances between the θ-estimates and the $\boldsymbol{\theta}$-points used to generate the data.

As was the case for the unidimensional IRT models, the precision of a test can be determined by summing the information available from each item. MIRT models have one additional complication. The sum of the information estimates must be for the same direction in the $\boldsymbol{\theta}$-space. Figure 5.11 shows the information surfaces in two dimensions for three different directions the space for the items described in Table 5.2. The leftmost figure shows the information from the 20 item test in a direction parallel to the θ_1 axis. The rightmost figure shows the information in a direction parallel to the θ_2 axis. The middle figure shows the information at a 45° angle to both axes. Note that the vertical axis is not the same for the three figures because there is more information in the center figure than for the ones on either side.

Beyond the fact that the high points of the three figures are in different places in the $\boldsymbol{\theta}$-space, it is important to recognize that the shapes of the test information surfaces are different. For estimating θ_1, the information surface shows that the most information is available between 0 and -2 on the θ_1 scale. The estimation of θ_2 is most accurate along a diagonal from $(-4, 4)$ to $(4, -2)$. The best estimation of an equally weighted combination of θ_1 and θ_2 is at about $(-2, 2)$ in the $\boldsymbol{\theta}$-space. These plots of the information surfaces do not provide a very clear picture of the information provided by the test because of the need to simultaneously attend to

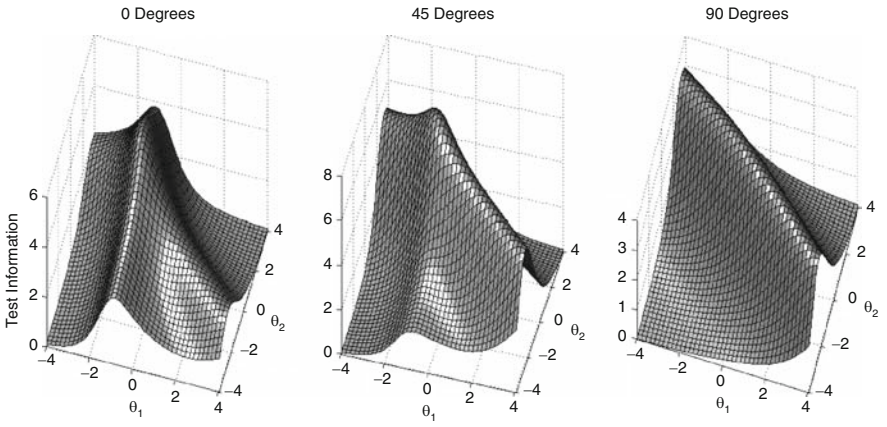

Fig. 5.11 Information surfaces for the test composed of items in Table 5.2 in three directions in the $\boldsymbol{\theta}$-space

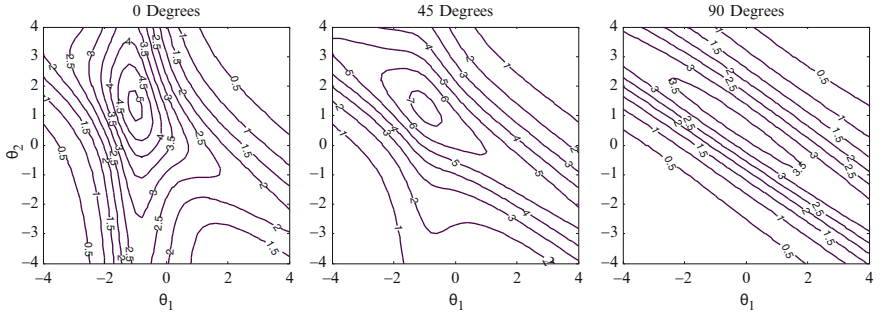

Fig. 5.12 Contour plots of information for the test composed of items in Table 5.1 in three directions in the $\boldsymbol{\theta}$-space

the features of multiple surfaces. Other alternative ways of to show the information available to estimate the θ-coordinates are to present the information as contour plots or the clam shell plots shown in Fig. 5.7.

The contour plots of the information for the same 20-item test are shown in Fig. 5.12. The contour plots make the location of the areas of highest information more evident and they also make it easier to determine which direction provides the most information. In this case, the 45° direction yields peak information over 7 and the other two directions are lower in the most informative areas. It is also clear that the test is most informative in all three directions near $(-1, 1.5)$ in the $\boldsymbol{\theta}$-space.

The clam shell plot of the test information for the same 20-item test is given in Fig. 5.13. In this case, the total test information is shown by the length of the line at each point in the direction indicated by the line. At the bottom of the plot is a line segment with the value 7.33 next to it. The line segment shows the length of line in the plot that represents 7.33 units of information. The clam shell plot shows the same pattern of information as the other plots, but now it is all included in one plot.

All of these representations of the information provided by a test are only practical when the functioning of the test items can be represented in a two-dimensional coordinate system. Although this limits the applicability of the graphic procedures for representing information, these examples still show that the amount of information provided by a test is a complex function of the location in the $\boldsymbol{\theta}$-space. There are portions of the space where estimation of an examinee's location would be very poor. For example, at $(-3.5, -.5)$ the information is almost zero for the test modeled here. At $(-1, -2.5)$, there is a small amount of information, but it is mainly useful for estimating θ_1. The test provides the most information in a variety of directions slightly below 0 on θ_1 and slightly above 0 on θ_2.

The complexity of the form of the test information surface suggests that it would be useful to search the $\boldsymbol{\theta}$-space to determine where the information is greatest and in what direction at that point. At the time of the writing of this book, there was no convenient procedure for finding the location and direction of maximum information for a test analyzed with several dimensions. With the current generation of

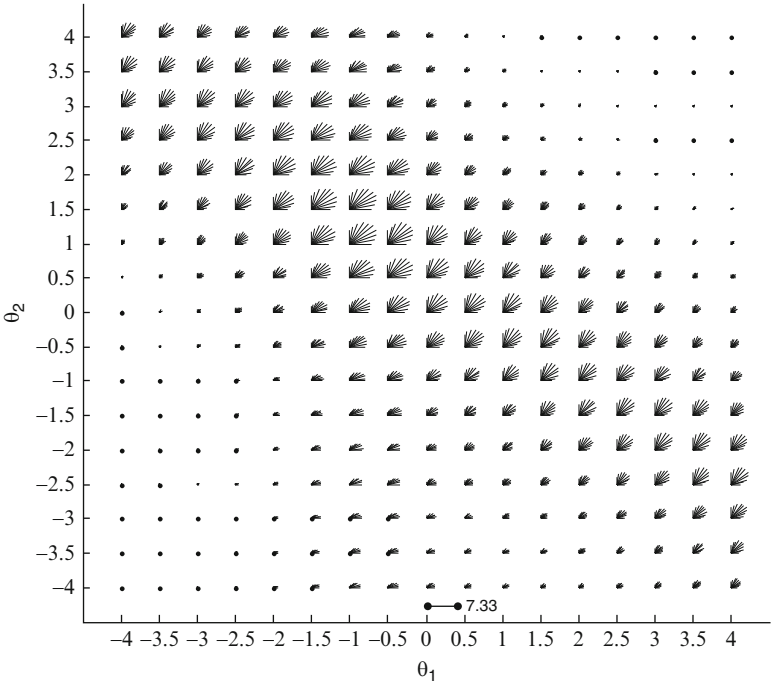

Fig. 5.13 Clam shell plot for the test information from the 20 items in Table 5.1

computers, it seems practical to search a grid of points in the space for the maximum information in each dimension, but software for that process has not been developed.

Another way to represent test results using MIRT is by tracking the multidimensional meaning of other types of scores computed from the item score vectors. For example, even tests that have very complex content structures often report scores that are based on the sum of item scores. MIRT procedures can be used to determine the meaning of this supposedly simple score type. The *centroid* plot is a way to show how the meaning of a single score derived from a test that requires multiple dimensions for successful performance changes with the level of the score. This form of representation was first used to check the multidimensional parallelism of test scores from different test forms (Reckase et al. 1988; Davey et al. 1989).

Centroid plots are formed by sorting the reporting scores into order from low to high. Then the sorted scores are divided into groups of equal size. For each group, the mean and standard deviation of the elements of the corresponding $\boldsymbol{\theta}$-vectors are computed. The resulting mean vectors are called group centroids. The standard deviations of the elements of the $\boldsymbol{\theta}$-vectors are used to put error ellipses around the centroids. One way to specify the error ellipses is to have the axes of the ellipses be equal to two times the standard error of the means for each dimension of the $\boldsymbol{\theta}$-vector.

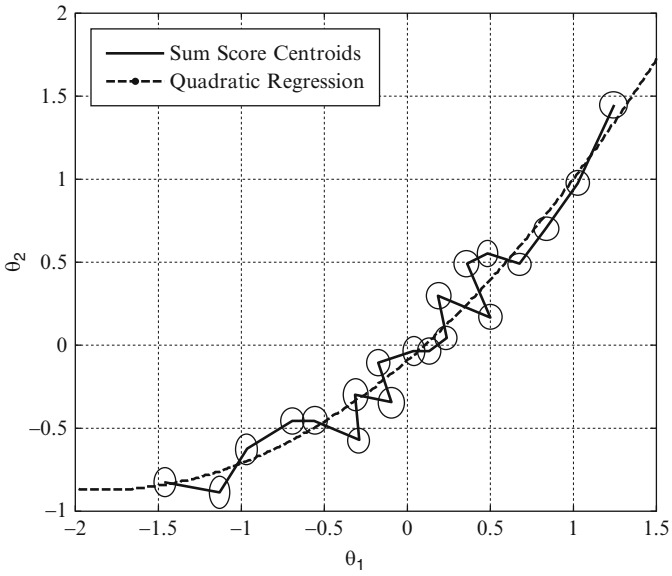

Fig. 5.14 Centroid plot based on number-correct scores for the test using items in Table 5.1

Figure 5.14 shows the centroid plot for the test composed of the items in Table 5.2 and a standard bivariate normal distribution of $\boldsymbol{\theta}$-coordinates with the identity matrix as the covariance matrix. The solid line on the graph connects each of the group centroids. Twenty groups were used in this case with a sample size of 100 for each group. The ellipses are located at each centroid and their axes are defined by the standard error of the mean of the corresponding coordinate. Ellipses that would overlap if they were on the same line indicate that the centroids are not significantly different from each other on that dimension. The plot also includes the quadratic regression line for the centroids showing that the centroids shift in the direction of change from the lower score levels to the higher score levels. The lower level centroids change in location mainly along θ_1. Higher centroids shift more along θ_2.

If θ_1 were aligned with a construct like arithmetic computation and θ_2 were aligned with a construct like problem solving, these results would indicate that differences in number-correct scores near the bottom end of the score scale are due to differences in arithmetic computation skills. Differences in problem solving skills have little effect at that level. At the high end of the score scale, problem solving skills are a major component in differences in the scores.

Centroid plots can be produced for any reported score for the test. For example, the UIRT θ-estimates could be used in place of the number-correct scores in the previous example. Centroid plots can also be produced for tests that are modeled in three dimensions. In that case, the mean vectors for each group define points in three-space and the standard errors of the means are used as the axes of an ellipsoid. Figure 5.15 shows a three-dimensional centroid plot using the number-correct score as the unidimensional variable.

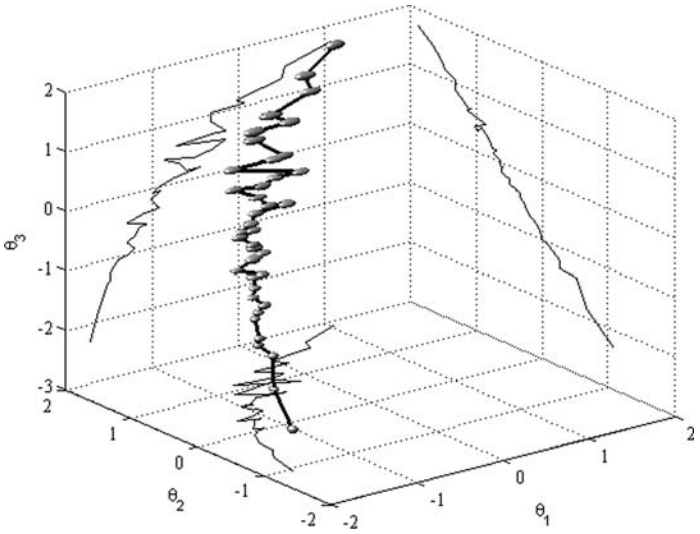

Fig. 5.15 Three-dimensional centroid plot with standard error of the mean ellipsoids and projections onto θ-planes

The plot in Fig. 5.15 is based on an analysis of 4,000 cases divided into 40 groups of 100 using the number-correct score for the test. The test is composed of 70 dichotomously-scored test items. The error ellipsoids for each centroid are fairly small indicating that the successive centroids are usually significantly different from each other. The plot also contains the projections of the lines connecting the centroids onto the side panels of the plot – the planes formed by pairs of θs. The projection onto the θ_2, θ_3 plane is close to linear while the projections onto the other planes are nonlinear. A two-dimensional analysis of this test based on θ_2 and θ_3 would indicate that the meaning of the differences on the number-correct score scale is the same over the range of the scale, but the full analysis of the data indicate that the meaning of differences on the number-correct score scale change with the level of performance on the scale.

5.4 Summary and Conclusions

This chapter provides ways for representing the characteristics of test items and tests when MIRT models are used to describe the interaction between persons and test items. Some of the statistical indices and graphical representations show the usual IRT item characteristics of difficulty and discrimination. Because MIRT models consider the functioning of items in a multidimensional space, there is also a need to indicate the direction in the space that the item provides the best discrimination. This is done by representing an item as an arrow or vector in the multidimensional space.

This type of representation shows the direction of greatest discrimination and helps show the relationships among items and the connection between unidimensional and multidimensional models.

The chapter also presents methods for describing the information provided by items and tests that can be used for estimating the location of a person in the $\boldsymbol{\theta}$-space. The information about a person's location in the space is shown to differ by the direction taken from that location. Items and tests are not equally good at differentiating individuals that differ from each other along different coordinate axes. The concept of information will prove useful in later chapters when test design issues are considered.

Finally, the chapter describes the relationship between unidimensional estimates of a construct and multidimensional representations. The results from unidimensional IRT models are shown to be weighted combinations of the estimated coordinates from the MIRT solution. The weights are related to the discrimination parameters for the items. Centroid plots are used to show that the meaning of reported scores may be different at different points along the score scale. These multidimensional representations were designed to help develop greater conceptual understanding of the way MIRT models represent test items and clarify the meaning of the results of the application of unidimensional models to multidimensional data.

5.5 Exercises

1. A set of test items require three coordinate axes to represent the interaction between the examinees and the test items. The parameters for two of the test items are provided in the following table.

Item number	a_1	a_2	a_3	d
1	1.2	.5	.4	−1
2	.35	.9	.9	1

For each item, compute the distance from the origin of the space to the point of steepest slope in the space. Which of the two items would be easier for a population of examinees who are distributed as a standard bivariate normal distribution with $\rho = 0$? What is the distance along the coordinate axes from the origin of the space to the point of steepest slope in that direction? Determine that distance for each item along each coordinate axes.

2. A person is located at a distance of 2 units from the origin of a three-dimensional space along a line with the following angles with each of the coordinate axes: [33.2 63.4 71.6]. The angles are in degrees with axes θ_1, θ_2, and θ_3, respectively, and the person has positive coordinates on each axis. Give the person's location as a $1 \times m$ vector of $\boldsymbol{\theta}$ coordinates.

3. Draw the arrow representation for the two items described by the parameters in Exercise 1.

4. Figure 5.14 is a centroid plot for the scores from a test that shows differences mostly along θ_1 for low scoring examinees and differences with more of a θ_2 component for high scoring examinees. Describe how you would select items for a test using the items' MIRT parameters so that both low and high scores show differences along θ_2 and mid-range scores show differences along θ_1. Give your reasons for the approach to selecting test items.