# 9
# Functional Models for Test Items

## 9.1 Introduction

After our bank accounts and our taxes, it is hard to imagine data playing a more central role in our lives than the examinations, opinion surveys, attitude questionnaires, and psychological scales administered to ourselves, our children, and our students. These data may not on first impression appear to be functional, but we show that functional data analysis can reveal how both test takers and test items perform in test situations. To provide a concrete frame of reference, we look at the responses of 5000 examinees to 60 items in a test of mathematics achievement developed by the American College Testing Program. We apply functional principal components analysis to explore variation across test items, and we check the fairness of certain items by comparing male and female performance. Finally, we use a functional property of these data to develop a useful new way of describing the performance of individual examinees.

Let us assume that each of $n$ items is given to each of $N$ examinees, and that each item is answered either correctly or incorrectly. We record each response with a value of 1 if examinee $j$ answers item $i$ correctly, and 0 otherwise. We want to use these data, crude as they may seem, to provide a reasonable answer to the question, "What is the probability $P_{ij}$ that examinee $j$ gets item $i$ right?"

Since we have only a single 0/1 datum to estimate $P_{ij}$, we obviously need to make some simplifying assumptions. We can take advantage of the fact that exam performances are not really all that unique; given this many

examinees, an arbitrary examinee $j$ is likely to have lots of "neighbors" in the sense of other examinees who get about the same number of right and wrong answers. Moreover, we will likely see that they even distribute these answers in a roughly similar manner. To a first approximation, poorly performing examinees will tend to get only the same easy items right, and strong examinees will fail only the same small subset of extremely hard items. Thus, we can pool information across similar examinees if we can propose a reasonable way of defining "similar."

## 9.2 The ability space curve

Figure 9.1 captures an idea that underlies almost all models for test data. We have plotted estimates of these right answer probabilities $P_{ij}$ for three test items on the ACT exam. Using a techique outlined below, these probabilities were estimated for 21 prototypical examinees, selected across the whole range of ability. Note that these are not actual candidates; rather, the observed data are used to obtain estimates of the probabilities of success for various items as the ability of the candidate varies in some way. Items 1, 9, and 59 were selected for Figure 9.1 because they are, respectively, low, medium, and high in difficulty. We can see that most of the 21 examinee points are high along the Item 1 axis, indicating that Item 1 is easy. Item 59's difficulty is demonstrated by the fact that most points are low along the corresponding axis, and because many points are in the middle of the range on the Item 9 axis, that item is somewhere between these two in difficulty.

The points corresponding to examinees fall along a curve. At the near end in Figure 9.1 are the poor students who pass all three items with probabilities near 0, and at the far end are those who rejoice in near certainty of passing all three. We use the term *space curve* to refer to a curve like this in a space of three or more dimensions. Of course, Figure 9.1 is only an incomplete picture; what we really have in mind is the space curve within 60-dimensional space, the coordinates of which are the probabilities of success on each of the 60 items. The *smoothness* of this space curve, or its continuum character, reflects a belief that probabilities of success will change smoothly as we change ability. Now of course there is such a thing as sudden insight, but the data collected by large testing agencies administering examinations to millions of people a year supports this assumption of a steady change in probability, at least for answers to multiple choice exam questions and for most examinees.

Our usual practice of summarizing test performance by a single score, such as number correct, also reflects these notions of unidimensionality and smoothness. We consider examinees as tending to vary in essentially one way that we refer to as low-to-high ability. When we group together ex-

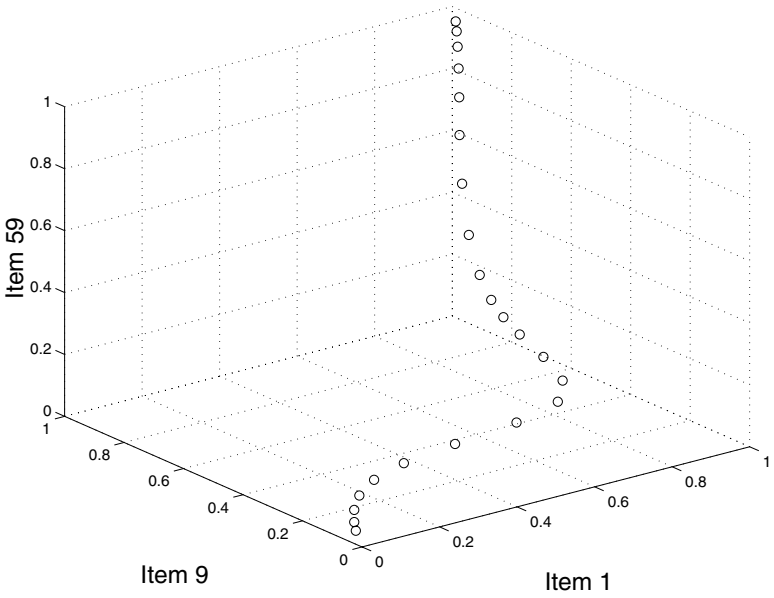


Figure 9.1. Each circle plots the three probabilities of success on items 1, 9, and 56 in the ACT math test for an examinee. The nearest 3 points are for examinees likely to fail all three items, and the far 3 points are examinees likely to succeed on all three. These 21 points fall along a smooth space curve within the unit cube.

aminees with the same test score, we expect to find that their patterns of right and wrong answers are not all that different. We also find that, as we move between nearby scores, the changes in these patterns are comparatively small. Indeed, tests are designed this way, by selecting items we know in advance will be easy, average, or hard. In short, if you are an average student taking a well-designed test, you and most other average students will fail the hard items, get the easy ones right, and differ from each other mostly in terms of the items that match your ability.

Thus, a plausible way to define "similar" for pairs of examinees is in terms of small differences in test scores. Two examinees have performances in the same "neighborhood" if their test scores are close together. We refine this notion later, but this seems like a reasonable place to start.

Any space curve can be defined by letting the coordinates of points on the curve be functions of a single variable. Consider, for example, a set of points in 3-D with coordinate values $X_i, Y_i$, and $Z_i$, and let these coordinate values be defined in terms of variable $z$ by the equations

$$\begin{aligned} X_i &= \sin(\pi z_i) \\ Y_i &= \cos(\pi z_i) \\ Z_i &= z_i. \end{aligned} \tag{9.1}$$

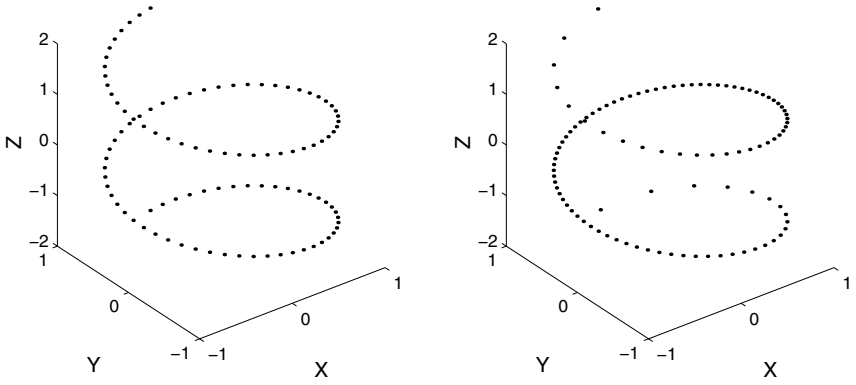


Figure 9.2. The locations of the points on the spiral in the left panel are determined by equations (9.1) for 101 equally-spaced values of $z$ between -2 and 2. In the right panel the points are determined by values of $z$ having a normal distribution.

Then the left panel of Figure 9.2 shows what happens if we let the variable $z_i$ take on 101 equally spaced values between $-2$ and 2. The variable $z$ is called the *charting variable.*

What if we made the values of $z$ have values at equal percentage points of a normal distribution within these limits? The result is in the right panel of Figure 9.2. Although the spacings between points have changed, the shape of the spiral has not. From this example, we can infer that the shape of a space curve will not change if we make any smooth order-preserving transformation of the variable $z$. This principle explains why we can have many different mapping systems for charting out the surface of the earth; the earth is the same whichever we use, but particular choices are more convenient for some purposes than others.

Let us therefore define examinee $j$'s position on the test performance curve in Figure 9.1 by the value $\theta_j$ of some charting variable $\theta$. Then what Figure 9.1 displays, and what is redisplayed in Figure 9.3, are the functions $P_i(\theta)$ indicating how probability of success on item $i$ varies over values of variable $\theta$. It seems reasonable to call $\theta$ a measure in some sense of "ability" or "proficiency," and it is referred to by psychologists as the *latent trait* underlying performance on the exam. The functions $P_i(\theta)$ are called *item response functions* or *item characteristic curves.*

However, our spiral example shows us that there is no unique way to define the variable $\theta$ that maps out the space curve. Psychometricians usually resolve this ambiguity by fiat by imposing the restriction that the values of $\theta$ in the population of examinees have a standard normal distribution, along the lines of the right panel of Figure 9.2. This choice is arbitrary, but it does reflect the long-standing assumption, or perhaps tradition, that

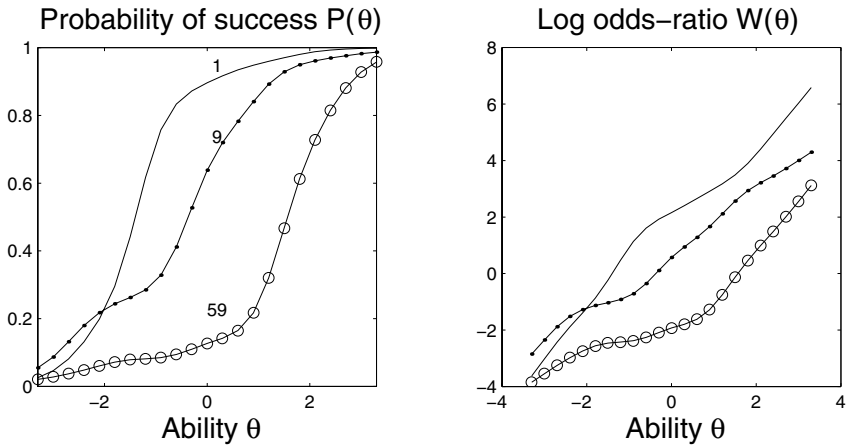


Figure 9.3. The three items displayed in Figure 9.1 are plotted in the left panel as functions $P_i(\theta)$ of the latent variable $\theta$. The right panel contains the plots of the corresponding log odds-ratio functions $W_i(\theta)$

ability has a roughly normal distribution. The classic example is IQ as a measure of intellectual ability. We will return to this issue later and propose an alternative variable that has some useful properties.

## 9.3 Estimating item response functions

Probability functions such as $P_i(\theta)$ present special computational challenges because they are constrained to take values only between 0 and 1. We can deal with this constraint by applying a suitable transformation, and a convenient reformulation of $P_i(\theta)$ is

$$P_i(\theta) = \frac{\exp[W_i(\theta)]}{1 + \exp[W_i(\theta)]} , \qquad W_i(\theta) = \log \frac{P_i(\theta)}{1 - P_i(\theta)} . \tag{9.2}$$

Values of $W_i(\theta)$ near 0 correspond to success probabilities in the vicinity of 0.5, large negative $W$s to very low $P$s, and large positive $W$s to near certainty of success. The function $W_i(\theta)$ is called the *log odds-ratio* function, and there are no constraints on its value.

The simple linear model

$$W_i(\theta) = a_i(\theta - b_i) \tag{9.3}$$

is one of the standard parametric models in psychometric theory, the two-parameter logistic model, or *2PL model* among those in the trade. Parameter $b_i$ of this model is called the *difficulty* of the item and captures the location of the log odds-ratio function, by specifying the value for which $P_i(\theta) = \frac{1}{2}$. The slope parameter $a_i$ is called the *discriminability* of the item,

and is an index of how well the test item distinguishes between test takers as $\theta$ varies. Although the curves $W_i(\theta)$ that we estimate for this test will usually be more complex in shape than this, these two qualities of location and slope are fundamental descriptors of item performance.

In practice, the 2PL model is too simple because for most multiple choice tests even the weakest examinees can achieve a positive success rate merely by guessing. Consequently, the industry standard model is the three-parameter logistic model or *3PL model*, which uses an additional parameter $c_i$ indicating this low-ability success probability, and has the structure,

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \ . \tag{9.4}$$

See Lord (1980) for a review of modern test theory and a wide range of applications of this model.

How do we estimate these log-odds functions $W_i(\theta)$ for each item, not knowing in advance what the independent variable values $\theta_j$ are for each examinee? The EM algorithm (Dempster, Laird, and Rubin, 1977) is used, in which $\theta_j$ is treated as if it were a missing datum. The EM algorithm proceeds by alternating between a phase called the E-step in which the item response functions are assumed known and likelihood is averaged over possible values of $\theta$, and the M-step in which the $\theta_j$s are assumed available for a small number of prototypical examinees and the functions $W_i(\theta)$ are estimated.

We achieved much more flexibility than in (9.3) or in (9.4) by expanding $W_i(\theta)$ in terms of 11 B-spline basis functions using equally spaced knots. We used a penalized EM algorithm, which maximizes the likelihood but also imposes a certain amount of smoothness on these estimated functions by using a roughness penalty based on the log odds-ratio. Details are found in Rossi, Wang, and Ramsay (2002).

## 9.4 PCA of log odds-ratio functions

Let us assume that the item response functions $P_i(\theta)$ and their log-odds equivalents $W_i(\theta)$ have been estimated to our satisfaction. We now want to explore how these functions vary from item to item.

Functional principal components analysis can reveal interesting aspects of the variation among these items. Because they are unconstrained, we apply PCA to the log odds-ratio functions instead of the probability functions. In this section we focus attention on functions estimated from the 2115 male candidates. The first four principal components of the 60 log odds-ratio functions then account for 96% of the variation; although there are quite a large number of test items, their characteristics are captured essentially completely by variability in four dimensions. Figure 9.4 shows

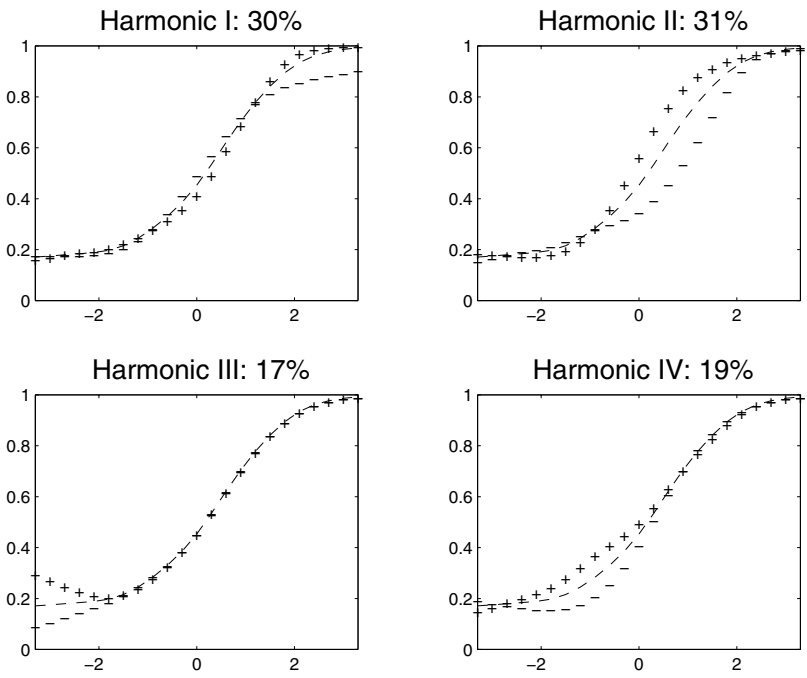


Figure 9.4. Each panel displays a varimax-rotated principal component of the variation among the log odds-ratio functions $W_i(\theta)$ estimated for the male candidates. A small multiple of each component is added (+) and subtracted (-) from the mean function, and the results transformed to probability functions, along the mean function. The percentages indicate percentages of variance accounted for, the total of which is 96%.

these four principal components after a varimax rotation to aid interpretation. These rotated components are displayed by adding and subtracting a small multiple of each component to the mean function $\bar{W}(\theta)$, and then back-transforming these perturbed means to their probability counterparts using (9.2).

These components can now be interpreted. Components I and III account for variation in characteristics of test items in the high and low ability ranges, respectively; components II and IV concentrate on variation over larger parts of the ability range, higher for component II and lower for component IV. An item with a high score on component I will be particularly good at sorting out very high ability students from others of moderately high ability, whereas if its score is low it will discriminate well among most of the population but will be found approximately of equal difficulty by all the very good students. Even the best students will not be certain of getting the item correct, a type of variation that the industry-standard 3PL model is unable to capture. However, we would be wise to

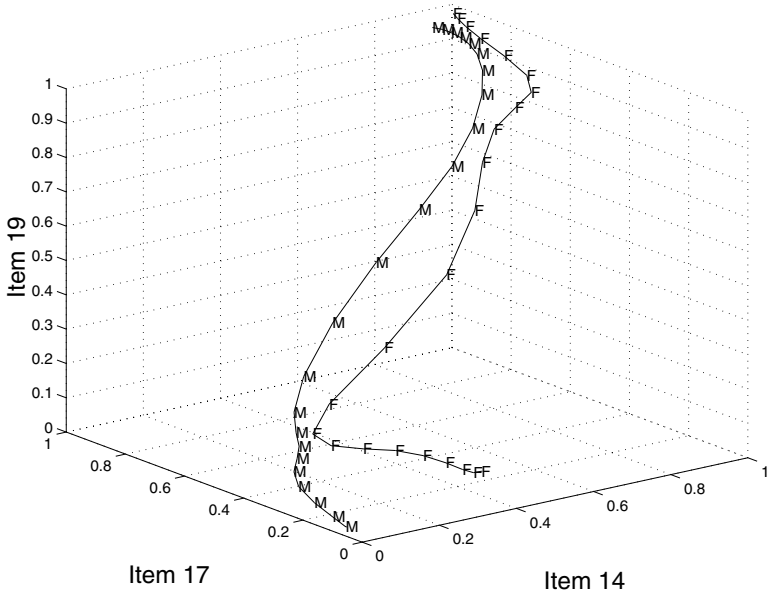


Figure 9.5. The space curves for items 14, 17, and 19 for men (M) and women (F).

remind ourselves that, even though the original data set is large, variation in the log-odds functions for extreme $\theta$ values is necessarily estimated by relatively small numbers of examinees, so conclusions for the extremes of the ability range should be treated with some caution.

An item with a high score on Component II would have a higher slope near the middle of the ability range and a lower slope for candidates with $\theta$ values approaching 2. Such an item gains local discriminability for average candidates at the expense of discriminability for the more able students. Similarly, Component IV quantifies a discriminability trade-off between average candidates and those with rather low abilities.

## 9.5 Do women and men perform differently on this test?

The ACT math test was taken by 2885 women and 2115 men. Figure 9.5 shows the space curves plotted in Figure 9.1 for both men and women for three different items. We see that performance on these three items evolves differently, and we may wish to investigate if there is something unusual about these three items.

We need a gold-standard summary of performance on the test such that for men and women having the same level on this summary, we can consider

that they are roughly equivalent in ability. We cannot use $\theta$ for this purpose, since we have forced this parameter to have a standard normal distribution within each group. In particular, the mean $\theta$ value is zero for each group, regardless of any way that the groups might differ in overall performance. The reason that the comparison is difficult is that there may be differences in the pattern of performance, not merely its level. What we need is a way of comparing the separately estimated $\theta$ values for women with $\theta$ values for men.

The performance measure that comes to mind immediately is the number of right answers as a function of $\theta$, and the expected value of this is

$$\tau(\theta) = \sum_{i}^{n} P_i(\theta) \ .$$

This *expected score* $\tau(\theta)$ measure of performance is often used by psychometricians to compare people in different groups.

However, we can propose some modifications of this idea. First, we might use the expected log odds-ratio, since in general it is wiser to take averages of unconstrained functions for the same reasons that we preferred to use PCA on the log odds-ratios. Once computed, we can back-transform this mean to the probability scale, and multiply it by the number of items to get what we might call a *fair score*. Second, we compute the expected value only using those items that do not appear to have gender differences in performance, so as to not contaminate our measure. In fact, only the three items plotted in Figure 9.5 appear to show much gender separation, so we use

$$\overline{W}(\theta) = (n-3)^{-1} \sum_{i \neq 14,17,19}^{n} W_i(\theta) \ ,$$

which we then back-transform to get our fair score

$$\tau(\theta) = \frac{\exp[\overline{W}(\theta)]}{1 + \exp[\overline{W}(\theta)]},$$

which we estimate separately for men and for women.

Figure 9.6 plots probabilities of success against fair score for men and women on items 17 and 19. Item 17 seems to favor men over most of the fair score range, and item 19 favors women. Item 14 is not plotted, but also favors men. These items exhibit what psychometricians call *differential item functioning*, abbreviated DIF. In the present context, it would probably make most sense in future tests to discard these three items altogether. An interesting question of a nonstatistical nature is to ask what is it that makes these mathematical items easier for one gender than another, when most are gender-neutral. It is especially interesting that the difference is not all in one direction.

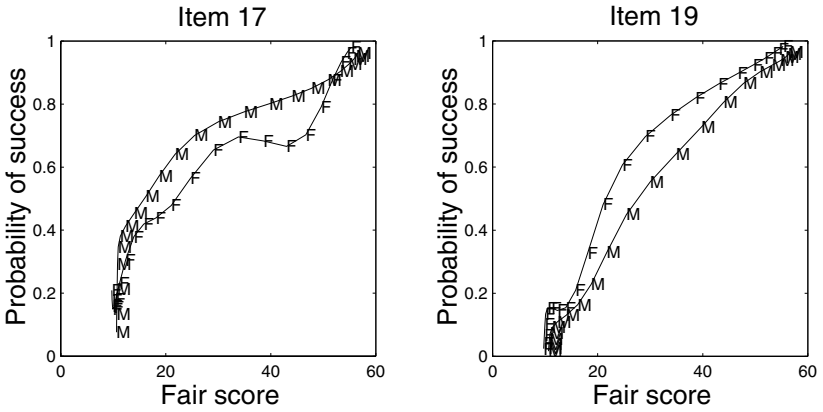


Figure 9.6. Probabilities of success for items 17 and 19 are plotted against a fair score that is a reasonable basis for equating ability of men and women.

## 9.6 A nonlatent trait: Arc length

In principle, there is nothing wrong with choosing the charting variable $\theta$ the way psychometricians do; the choice is arbitrary, and if one likes to think of ability as normally distributed, their choice is appealing. Unfortunately, users of test theory models, and some psychometricians as well, have tended to lose sight of the arbitrariness of the choice, and fall into thinking that the values $\theta_j$ measure ability in the same metric sense that the marks on a ruler measure length. It has been claimed, in fact, that this is one of the big arguments for using latent trait theory to model test performance.

Actually, there is a charting variable that really does have the metric properties that users and theorists would like to see, and is moreover not at all latent. This is *arc length*, $s$, the distance along the space curve determined by the simultaneous changes in probability as we move along the curve. We have already used arc length to advantage in Chapter 8 as a way of describing curves in two dimensions.

Arc length resists misinterpretation because small changes $\Delta s$ in distance along the curve really do have a meaning that does depend on our present position. Distances along the curve are directly related to the changes in probabilities of success for the test items. Like units of physical measurement, arc length differences can meaningfully be added and subtracted.

The values of arc length $s$ are computed by beginning with some arbitrary charting variable such as $\theta$, estimating the corresponding item response functions $P_i(\theta)$ and their derivatives $P_i'(\theta)$, and then computing arc length

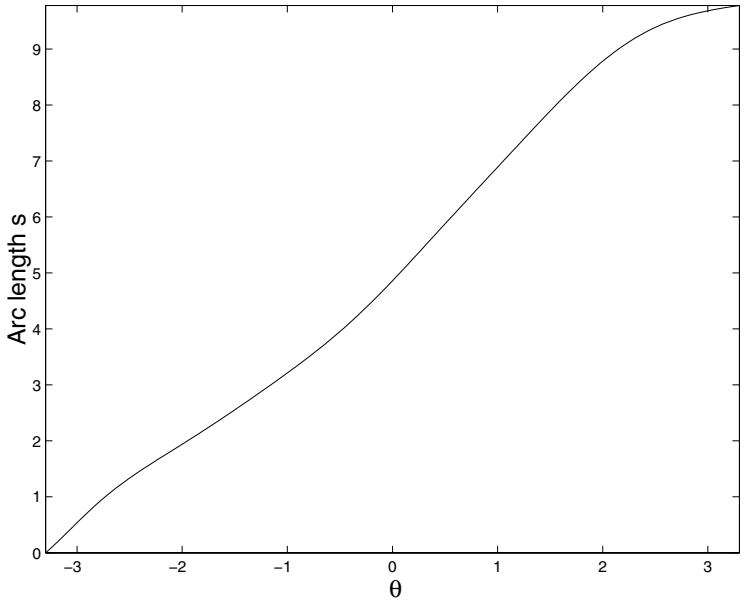


Figure 9.7. Arc length from a reference point, or the distance along the ability space curve, as a function of standard normal latent variable $\theta$.

$s(\theta)$ by the equation

$$s(\theta) = \int_{\theta_0}^{\theta} \left\{ \sum_i [P_i'(u)]^2 \right\}^{1/2} du. \tag{9.5}$$

In this equation $\theta_0$ is the lowest value of $\theta$ on the curve.

Arc length is called the *intrinsic metric* of the space curve, because its values do not depend on what kind of charting variable we use in (9.5). For the spiral in Figure 9.2, the 101 equally spaced values between 0 and $4\sqrt{2}$ are of equal arc distance along the curve.

For the male candidates in the math test, with the usual charting variable $\theta$ having a standard normal distribution, arc length $s(\theta)$ is displayed as a function of $\theta$ in Figure 9.7. We see that, in fact, the relationship is close to linear for all except the highest values of $\theta$. Therefore, in this context arc length does not represent any dramatic departure from the traditional $\theta$ measure. The reference point from which arc length is measured corresponds to the performance of the weakest examinee.

For purposes of communicating with a user community, we would not mislead anyone much by linearly rescaling arc length to have an upper limit of 100 while retaining the lower limit of 0. The metric properties of this rescaled measure would still hold. Alternatively, as the Educational Testing Service and other large testing agencies do, we can pick lower and upper

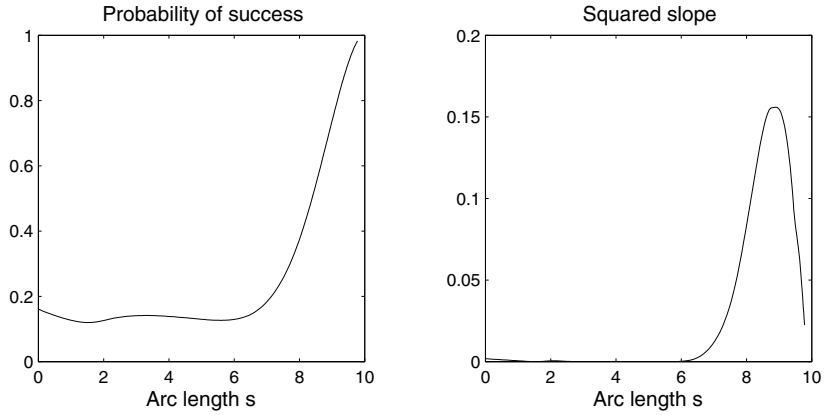


Figure 9.8. The left panel contains the item response function for item 56 as a function of arc length $s$, and the right panel contains its squared slope, a normalized measure of item quality. Only items 57 and 60 are this discriminating for high performance examinees.

fixed limits and rescale arc length to be within these limits. This would still be a metric measure of performance in the sense that differences can be added.

The elements $P_i'(s)$ of the tangent vector are the slopes of the item response functions at arc length $s$, and therefore measure the discriminability of the item. Arc length as a charting value has a useful property for assessing the quality of an item. Because we move at a steady speed along the curve as arc distance increases, the length of the tangent vector $\{P_1'(s), \ldots, P_n'(s)\}$ is exactly 1 when the curve is parameterized by arc length. Thus,

$$\sum_{i=1}^{n} \left( \frac{dP_i}{ds} \right)^2 = 1.$$

Since the squares of the discriminability estimates must sum to one, we can compare them across items by plotting $[P_i'(s)]^2$. The test items particularly contributing to discriminability will be different at different parts of the ability range.

For example, test developers find it hard to construct an item that discriminates well for examinees at the upper end of the ability continuum. Item 56 turns out to be such an item, and Figure 9.8 displays its item response function and its squared slope or discriminability as functions of arc length. The fact that the latter exceeds 0.15 and that the sum across all items of squared discriminability is 1 means that few items are this discriminating. In fact, only this and items 57 and 60 achieve any quality for high-end examinees.

We have highlighted items 56, 57, and 60 by considering the components of the tangent vector as functions of arc length. These results can be related

to the principal components analysis carried out above. The four lowest principal scores for Component II are for items 56, 57, 59, and 60. The items also have large negative scores on Component IV. Figure 9.4 and the discussion of the components in Section 9.4 indicate that items with negative scores on both these components will be best at sorting out able students from one another.

## 9.7 What have we seen?

Functional data analysis is not only a method for analyzing observed curves; it can also be applied to curves implied by and estimated from data that are not at all curvaceous at first sight. Any single test datum does not by itself provide a lot of information about the item success probability $P_{ij}$, but by making the strong simplifying assumption that these probabilities vary in a smooth one-dimensional way across examinees, we can estimate the ability space curve that this assumption implies.

Once we have chosen a charting variable $\theta$ to measure out positions along this space curve, we can also study the $n$ item response functions $P_i(\theta)$ as if they were a sample of observed functions. Actually, though, we are perhaps better off applying functional data analysis to the log odds-ratio functions $W_i(\theta)$, since these transformations of the item response functions have the unconstrained variation that we are used to seeing in directly observed curves. Principal components analysis seems like the ideal tool to study variations among these curves, and we found that the dimensionality of this variation was perhaps surprisingly small, and quite interpretable.

In the test item context, arc length is an attractive method of parameterizing ability. Arc length is not latent, may be less confusing to the practitioners of psychometrics, and offers an interesting new way of assessing item quality by plotting the square of the test discriminability function.

## 9.8 Notes and bibliography

To read more about modern test theory and its applications using parametric models, see Lord (1980) and the more classic Lord and Novick (1968). The EM algorithm was first applied to the estimation of parametric models in test theory by Bock and Aitkin (1981). Our use of the EM algorithm to estimate the functions $P_i(\theta)$ and $W_i(\theta)$ nonparametrically is based on theses by Wang (1993) and Rossi (2001), and are described in Rossi, Wang, and Ramsay (2002). The use of ideas from differential geometry to present nonparametric modern test theory comes from Ramsay (1995) and (1996a).