SOME LATENT TRAIT MODELS

17.1 Introduction

In this chapter we shall consider in detail several models of tests, some of which have been introduced more briefly above (Sections 15.6 and 16.1 through 16.5). We shall now describe these models in self-contained mathematical terms to prepare ourselves to examine them, subsequently, in relation to theories and applications of tests. These models have been developed primarily in connection with tests of various general or special abilities, although it has proved of interest to consider them also in relation to the study of other kinds of traits, such as attitudes. For convenience, we shall refer to the trait in question simply as "ability".

We consider here tests consisting of items each to be scored 0 or 1, with u_g as the generic symbol for the score on item g and with $\mathbf{v}' = (u_1, \ldots, u_g, \ldots, u_n)$ representing the set of scores, or the *response-pattern*, on a test of n items. This notation tacitly refers to scores of some one individual subject; when necessary, scores of a subject indexed a can be denoted more explicitly by $\mathbf{v}_a' = (u_{1a}, \ldots, u_{ga}, \ldots, u_{na})$.

Item scores u_g are related to an ability θ by functions that give the probability of each possible score on an item for a randomly selected examinee of given ability. These functions are

$$Q_g(\theta) = \text{Prob}\,(U_g = 0 \mid \theta)$$

and the *item characteristic curve* (*ICC*)

$$P_g(\theta) = \text{Prob}\,(U_g = 1 \mid \theta) = 1 - Q_g(\theta).$$

These formulas are conveniently combined in the probability distribution function of U_g:

$$f_g(u_g \mid \theta) \equiv \text{Prob}\,(U_g = u_g \mid \theta) = P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \equiv \begin{cases} P_g(\theta) & \text{if} \quad u_g = 1, \\ Q_g(\theta) & \text{if} \quad u_g = 0, \end{cases}$$

where f_g is defined in a persons or in a persons-by-replications space.

We note that the regression function of any item response u_g is identical with its item characteristic curve since

$$\mathscr{E}(U_g \mid \theta) = 1 \cdot f_g(1 \mid \theta) + 0 \cdot f_g(0 \mid \theta) = P_g(\theta).$$

Any item for which $P_g(\theta)$ has a constant value independent of θ is not an indicant (and hence *a fortiori* not a measure) of θ in the sense of Section 1.4. In most cases of interest here, we shall have $P_g(\theta)$ strictly increasing in θ, so that u_g will be an indicant and a measure of θ. We do not assume a probability distribution for θ in any part of the present treatment of this subject. For an extension of the theory which makes use of this assumption, the reader should see Birnbaum (1967).

These functions do not determine unequivocally the relation between an ability and a complete response pattern $\mathbf{v}' = (u_1, \ldots, u_n)$ unless they are supplemented in some definite way. The additional assumption found most useful in test theory and its applications, as well as the simplest assumption mathematically, is *local independence* (see Section 16.3). This assumption implies the mathematical condition of *statistical independence between responses* by a subject to different items; it is represented by the usual probability product form

$$\begin{aligned} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) &\equiv \text{Prob}\,(U_1 = u_1, \ldots, U_n = u_n \mid \theta) \\ &= \text{Prob}\,(U_1 = u_1 \mid \theta) \cdots \text{Prob}\,(U_n = u_n \mid \theta) \\ &= \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}. \end{aligned}$$

For example, the product form $\text{Prob}\,[(U_1, U_2) = (1, 1) \mid \theta] = P_1(\theta)P_2(\theta)$ represents the fact that any subject of ability θ gives independent responses to items 1 and 2; that is, that the probability $P_2(\theta)$ of his correctly answering item 2 is the same as the conditional probability of his correctly answering item 2, given that he has correctly answered item 1. The relations of this assumption to more general models and theories, in which several abilities are considered jointly, have already been discussed in Section 16.2.

One basic aspect of the questions of validity and empirical and theoretical content discussed in Chapter 1, as they apply to the models introduced here, may be illustrated conveniently at this point. Consider any item, and consider a series of groups of subjects in which each subject is assumed to have common ability. Suppose that the probabilities of correct responses to the item in the respective groups are p_i, where $0 \leq p_1 < p_2 < \cdots < p_m \leq 1$. Since we have mentioned all the empirically meaningful aspects of a model of a single item, we still remain free to choose arbitrarily a series of numbers θ_i,

$$-\infty < \theta_1 < \theta_2 < \cdots < \theta_m < \infty,$$

which we may call the *true ability scores* of the respective groups. The choice

of these numbers θ_i amounts to a choice of the specific form of an ICC function that shall represent the first item, since we *define* the function $P_1(\theta)$ as the correspondence between respective ability scores θ_i and values $P_i = P_1(\theta_i)$. Equivalently, given the numbers p_i, we can adopt *any* increasing function $P_1(\theta)$ as the ICC of the item: This choice associates an ability score θ_i, determined by $p_i = P_1(\theta_i)$, with the group of subjects scoring p_i.

These comments illustrate the fact that an essentially conventional element exists in the relations between ability levels θ and observable item responses. Once any specific strictly increasing form has been adopted for $P_1(\theta)$, for example, $P_1(\theta) = \Phi(2\theta - 1)$, the statement that a subject has ability $\theta = 2.1$ has empirical content and consequences in the contexts of models discussed here. For any second item (assuming local independence), the item characteristic curve $P_2(\theta)$ has a value at $\theta = 2.1$ which is estimable from empirical data in the same sense as is $P_1(2.1)$. Thus we are *not free* to adopt by definition *any* number as the value of $P_2(2.1)$. Similarly we are not free to adopt *any* assumption restricting even partially the possible functional forms of any other item characteristic curves $P_g(\theta)$, $g = 2, 3, \ldots, n$. This illustrates the fact that in general it is empirically meaningful (nontautological) to assume that any specific model, or even any class of models of partially restricted form, is valid in relation to a specified population of items. Therefore it is possibly false and hence is subject to empirical confirmation (or partial confirmation or disconfirmation). On the other hand, the assumption that any chosen *single* item has an item characteristic curve of a specified functional form $P_g(\theta)$ that depends on ability θ is, when considered *in isolation*, acceptable in principle as a definition of the ability scale of θ values and is not an empirical specification.

17.2 The Logistic Test Model

A function which very nearly coincides with the normal ogive model treated in Section 16.5, and which has advantages of mathematical convenience in several areas of application, is the logistic (cumulative) distribution function

$$\Psi(x) = e^x/(1 + e^x) \equiv 1/(1 + e^{-x}), \qquad -\infty < x < \infty. \tag{17.2.1}$$

The inverse function is $x = \log [\Psi/(1 - \Psi)]$. For simple descriptive purposes, any graph of a cumulative normal distribution function $\Phi(x)$ would serve equally well to illustrate this function, since it has been shown (Haley, 1952, p. 7) that

$$|\Phi(x) - \Psi[(1.7)x]| < 0.01 \qquad \text{for all } x. \tag{17.2.2}$$

We may state this relation in another way: The logistic cdf $\Psi(x)$ differs by less than 0.01, uniformly in x, from the normal cdf with mean zero and standard deviation 1.7; that is,

$$|\Phi(x/1.7) - \Psi(x)| < 0.01 \qquad \text{for all } x.$$

The probability density function (pdf) corresponding to the logistic cdf is

$$\psi(x) = e^{-x}/(1 + e^{-x})^2 \equiv \Psi(x)[1 - \Psi(x)] \equiv \tanh^{-1}(x). \qquad (17.2.3)$$

Berkson (1957) has given detailed tables of $\Psi(x)$ and $\psi(x)$. Of course, tables of the exponential function and of the hyperbolic tangent are also available, and hence direct computation of values of these functions is not difficult.

The *logistic test model* is determined by assuming that item characteristic curves have the form of a logistic cumulative distribution function:

$$P_g(\theta) = \Psi[DL_g(\theta)] \equiv [1 + e^{-DL_g(\theta)}]^{-1} = [1 + e^{-Da_g(\theta - b_g)}]^{-1}, \qquad (17.2.4)$$

where $L_g(\theta) = a_g(\theta - b_g)$, and $g = 1, 2, \ldots, n$. We have also

$$Q_g(\theta) = 1 - \Psi[DL_g(\theta)] \equiv [1 + e^{DL_g(\theta)}]^{-1},$$

$$P_g(\theta)/Q_g(\theta) = e^{DL_g(\theta)}, \qquad \text{and} \qquad \frac{\partial}{\partial\theta} P_g(\theta) = Da_g P_g(\theta) Q_g(\theta).$$

(Again, we do not interpret $P_g(\theta)$ here as a probability distribution function, even when it has the mathematical properties of one.) Here a_g and b_g are item parameters whose roles are generally the same as those of the item parameters in the normal ogive model because of the qualitative, and nearly exact quantitative, similarity between the models. The symbol D denotes a number that serves, at our convenience, as a unit scaling factor. To maximize agreement between quantitative details in the normal and logistic models, we can and usually shall take $D = 1.7$; then

$$P_g(\theta) = \Psi[1.7a_g(\theta - b_g)] \equiv (1 + e^{-1.7a_g(\theta - b_g)})^{-1}. \qquad (17.2.4a)$$

For notational convenience, however, we shall often write the logistic model using the symbol D for the number 1.7.

We may view the logistic form for an item characteristic curve as a mathematically convenient, close approximation to the classical normal form, introduced to help solve or to avoid some mathematical or theoretical problems that arise with the normal model. Or we may view it as the form of a test model that is of equal intrinsic interest and of very similar mathematical form. The important questions of the validity of such models in observational and theoretical contexts are discussed elsewhere (see Sections 16.1 and 17.10).

The probability distribution function of a response u_g in a logistic test model is

$$f_g(u_g \mid \theta) \equiv P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}$$

$$\equiv Q_g(\theta)[P_g(\theta)/Q_g(\theta)]^{u_g} \qquad (17.2.5)$$

$$= \frac{\exp[Da_g(\theta - b_g)u_g]}{1 + \exp[Da_g(\theta - b_g)]}, \qquad (17.2.6)$$

and, under the assumption of local independence, the probability distribution function of a response pattern $\mathbf{v}' = (u_1, \ldots, u_n)$ is

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \prod_{g=1}^{n} f_g(u_g \mid \theta) = \prod_{g=1}^{n} Q_g(\theta) \prod_{h=1}^{n} \exp\,[Da_h(\theta - b_h)u_h]$$

$$= \left[\prod_{g=1}^{n} Q_g(\theta)\right]\left[\exp\left(\theta D \sum_{g=1}^{n} a_g u_g\right)\right]\left[\exp\left(-D \sum_{g=1}^{n} a_g b_g u_g\right)\right]. \tag{17.2.7}$$

The principal features of mathematical simplicity that characterize the logistic test model are, as we shall see, implicit in this last form. In particular, "all the information about θ available in a response pattern $\mathbf{v}$" (in a sense to be specified) is given by the particular test score formula

$$x = x(\mathbf{v}) = \sum_{g=1}^{n} a_g u_g,$$

which does not depend on the difficulty parameters b_g. We may further illustrate the roles of item parameters and the properties of such a test score formula by considering an artificial example of the logistic test model. Let us take just four of the items whose parameters have the values represented in Fig. 16.5.1, namely, $g = 3, 4, 5$, and 6. (The same figure serves equally well here to illustrate either logistic or normal item characteristic curves.) We have $a_3 = 100$, $a_4 = 100$, $a_5 = 1$, and $a_6 = 1$. The test score is then

$$\begin{aligned} x &= 100y_3 + 100y_4 + y_5 + y_6 \\ &= 100(y_3 + y_4) + (y_5 + y_6). \end{aligned}$$

The possible values of x are just

0	1	2
100	101	102
200	201	202.

We see that the major part of this ordering of subjects' response patterns, which is represented by the rows of the preceding array, is determined by the heavily weighted responses to the informative items y_3 and y_4. The only role of the less informative items in this example is to give a finer ordering compatible with the initial rough ordering. This example is extreme: Typical tests one meets in practice have more items and less extreme variation in weights a_g. With more nearly typical tests, it is usually possible to reverse an ordering of two response patterns based only on responses to several items if all items are taken into account in a suitable weighted composite score.

Table 17.3.1

Standard deviation of sample item-test biserials*

Test	Number of items	Sample item-test biserials	
		Mean	Standard deviation
Listening Comprehension	50	0.51	0.12
English Structure	70	0.48	0.11
Vocabulary	60	0.55	0.09
Reading Comprehension	30	0.54	0.09
Writing Ability	60	0.44	0.11

* From an internal Educational Testing Service report (SR–66–80) prepared by Dr. Frances Swineford.

17.3 Other Models

If we assume a common value for the discriminating powers of the items, each $a_g = 1$, say, and take $D = 1$, we obtain the form

$$P_g(\theta) = \Psi(\theta - b_g) \equiv (1 + e^{b_g - \theta})^{-1}.$$

We can write

$$\theta^* = e^{\theta} \qquad \text{and} \qquad b_g^* = e^{b_g}$$

to denote, respectively, an ability parameter and an item difficulty parameter, each represented on a transformed scale. Then we have

$$P_g(\theta) \equiv P_g^*(\theta^*) = \left(1 + \frac{b_g^*}{\theta^*}\right)^{-1} = \frac{\theta^*}{b_g^*}\left(1 + \frac{\theta^*}{b_g^*}\right)^{-1}.$$

Rasch (1960) has developed the test model of this restricted logistic form. We see that this model is a special case of the logistic model in which all items have the same discriminating powers, and all items can vary only in their difficulties. Whenever this special logistic model holds, the considerable body of theoretical and practical methods developed by Rasch is applicable (see Chapter 21).

One very important question emerges at this point: Do the items in a test really differ from each other in discriminating power? This question is crucial to evaluating the validity of the models and methods of this and the following three chapters and to comparing these evaluations with evaluations of the validity of the simpler models and methods of Chapter 21. Some available item analysis data suggest an affirmative answer for multiple-choice paper and pencil tests. These data, which are represented in Table 17.3.1, are based on a sample of 3805 examinees. The table shows the mean and standard deviation of the sample biserial correlation between item score and test score for each of

five different tests. If the true biserial correlation is 0.50 in a normal population of this size ($N = 3805$), then the standard error of a biserial correlation from a sample of this size will only be from about 0.016 to about 0.019, depending on the item difficulty. (An approximate formula appears in McNemar, 1962, Eq. 12.3.) Since the standard deviations in this particular sample are at least five times as large as this standard error, it is clear that the variation found here among item-test biserials is almost entirely due to real differences among the item discriminating power parameters. In this sample we find that even if we disregard the five percent of the items with the highest and the five percent with the lowest discriminating power parameters, we still have a range from about 0.31 to about 0.67. Since item-test biserials approximate item-ability biserials, whose close relation to the slope of the item characteristic curve was discussed rather fully in Section 16.10, it is clear that the item characteristic curves of the items in Table 17.3.1 differ from each other by more than a mere translation (change of origin).

If $\theta > b_g$, then for any fixed values of b_g and θ,

$$\Phi[a_g(\theta - b_g)] \quad \text{and} \quad \Psi[Da_g(\theta - b_g)]$$

both increase to 1 as a_g increases; and if $\theta < b_g$, then both decrease to 0 as a_g increases. We may represent these limiting values formally as

$$\Phi[\infty(\theta - b_g)] = \Psi[\infty(\theta - b_g)] \equiv \begin{cases} 1 & \text{if} \quad \theta > b_g, \\ 0 & \text{if} \quad \theta < b_g, \end{cases}$$

since

$$(\theta - b_g)\infty \equiv \begin{cases} +\infty & \text{if} \quad \theta > b_g, \\ -\infty & \text{if} \quad \theta < b_g. \end{cases}$$

For convenience, we can give the value 1 to the otherwise undefined symbols $\Phi(\infty \cdot 0)$, $\Psi(\infty \cdot 0)$. Then we may define an item characteristic curve by

$$P_g(\theta) = \Phi[\infty(\theta - b_g)] \quad \text{or} \quad \Psi[\infty(\theta - b_g)].$$

These may be considered extreme, limiting cases of ICCs within the normal ogive and the logistic test models. Such ICCs do not have the property, generally assumed above, of increasing continuously and strictly as θ increases. Each is characterized fully by a single difficulty parameter b_g; for abilities $\theta < b_g$ it has the value zero, and at this ability level it increases discontinuously to unity. These curves may be regarded as representing items whose responses y_g are error-free indicants of abilities, in the sense that taking $y_g = 1$ as indicating $\theta \geq b_g$ and $y_g = 0$ as indicating $\theta < b_g$ entails probability zero of erroneous indications for each possible value of θ. It may be said that ICCs of this extreme form give "perfect scaling", since an ordering of subjects' abilities θ on the basis of any test consisting of such items is error-free (with probability 1, or certainty). Such items are basic to the scaling methods and the theory

developed by Guttman (1950), particularly in connection with scaling of latent traits θ representing attitudes.

Lazarsfeld has developed several classes of latent trait models, but primarily for the investigation of attitudes rather than abilities. One of these may conveniently be described here. If

$$P_g(\theta) = a_g(\theta - b_g), \qquad g = 1, \ldots, n,$$

where θ is restricted to an interval on which all values of $P_g(\theta)$ lie between 0 and 1, we have the *linear model.* Despite quantitative differences, here, as in other models described above, the item parameter a_g represents discriminating power in the sense of rate of change of $P_g(\theta)$ with respect to θ, and b_g locates the part of the θ scale where the item is effective. In Chapter 24, we shall present several other models developed by Lazarsfeld.

Methodological problems related to these models are discussed briefly by Torgerson (1958, Ch. 13), who gives references to basic papers and subsequent work. A later discussion is that of Lazarsfeld (1959).

Even subjects of very low ability will sometimes give correct responses to multiple-choice items, just by chance. One model for such items has been suggested by a highly schematized psychological hypothesis. This model assumes that if an examinee has ability θ, then the probability that he will *know* the correct answer is given by a normal ogive function $\Phi[a_g(\theta - b_g)]$ of exactly the kind considered in Section 16.5; it further assumes that if he does not know it he will guess, and, with probability c_g, will guess correctly. It follows from these assumptions that the probability of an incorrect response is

$$Q_g(\theta) = \{1 - \Phi[a_g(\theta - b_g)]\}(1 - c_g),$$

and that the probability of a correct response is the item characteristic curve

$$P_g(\theta) = c_g + (1 - c_g)\Phi[a_g(\theta - b_g)]. \tag{17.3.1}$$

The psychological hypothesis implicit here has been mentioned primarily to point up a mathematical feature of this form; the empirical validity of this form is not dependent on this psychological hypothesis. This model is possibly more reasonable than the random-guessing models discussed in Sections 14.3 and 14.5.

The function (17.3.1) approaches its minimum c_g as θ decreases. Its graph is that of a normal ogive curve except that the range of ordinates 0 to 1 is replaced by the range c_g to 1. If one of five multiple-choice alternatives were chosen at random whenever guessing occurred, we would have $c_g = \frac{1}{5}$, as in Fig. 17.3.1, where the other item parameters are equal to those in Fig. 16.5.1. Each of the general illustrative comments above concerning the item parameters a_g and b_g of normal ogive models can be adapted to apply to their roles in these models.

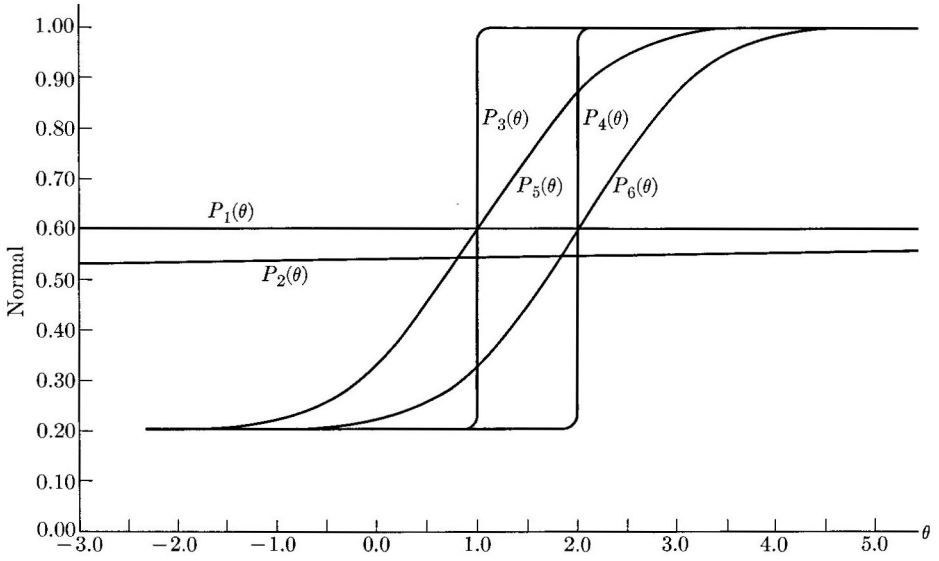

FIG. 17.3.1. Three-parameter normal ogive and logistic item characteristic curves.

Similarly with the logistic model, we may take account of guessing probabilities by using modified item characteristic curves, which here assume the form

$$P_g(\theta) = c_g + (1 - c_g)\Psi[Da_g(\theta - b_g)],$$

which Fig. 17.3.1 serves to illustrate. More detailed consideration of the roles of item parameters in such models is given below.

17.4 The Test as a Measuring Instrument: Examples of Classification and Estimation of Ability Levels by Use of Test Scores

We shall find it useful to consider the mathematical model of a test as having dual but related purposes. One purpose is to determine the value θ of an examinee's ability with adequate precision; the second is to classify an examinee into ability categories with adequately small probabilities of misclassification. We shall present brief descriptions of some estimation and classification methods based on test scores. These will illustrate some of the applications of the theory that we shall develop. Each of the simplifying assumptions or restrictions made here will require critical reconsideration later.

We shall consider a model of a test, represented by a specified probability function

$$\text{Prob}\,[\mathbf{V}' = (u_1, \ldots, u_n) \mid \theta],$$

possibly having one of the forms described above, in which the ability θ is the

only unknown parameter. We shall adopt a specified test score formula $x = x(\mathbf{v}) \equiv x(u_1, \ldots, u_n)$. These two functions determine the cdf of the test score:

$$F(x \mid \theta) = \text{Prob}\,[X(\mathbf{V}) \leq x \mid \theta] \equiv \sum_{X(V) \leq x} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta). \quad (17.4.1)$$

Numerical determinations of $F(x \mid \theta)$ for a number of score formulas and tests will be illustrated. In the simplest case, that of items having identical characteristic curves

$$P \equiv P(\theta) \qquad \text{and} \qquad x = \sum_{g=1}^{n} u_g,$$

the cdf, $F(x \mid \theta)$, is just the binomial cdf for n trials with parameter $P(\theta)$:

$$F(x \mid \theta) = \sum_{k=0}^{x} \binom{n}{k} P^k Q^{n-k}, \qquad x = 0, 1, \ldots, n. \quad (17.4.2)$$

Local independence is assumed here.

In most cases of interest, the magnitudes of discontinuities in $F(x \mid \theta)$ (that is, the probabilities of the individual possible values of x) will all be small for each θ, usually of the order of several percent or less. For many theoretical and practical purposes, it is convenient to treat $F(x \mid \theta)$ as continuous in x for each fixed θ, and also it is sometimes convenient to employ specific continuous functions of x as working approximations subject to appropriate bounds or independent checks on the approximations entailed. For illustrative simplicity, we treat $F(x \mid \theta)$ in this section as continuous and assume that for each fixed θ, it is strictly increasing from 0 to 1 with x. In the preceding binomial example, the convenient approximation is the usual one by the normal cdf (see, for example, Lindgren, 1962, p. 149):

$$F(x \mid \theta) \equiv \sum_{k=0}^{x} \binom{n}{k} P^k Q^{n-k} \doteq \Phi\left[\frac{x + \frac{1}{2} - nP}{(nPQ)^{1/2}}\right], \quad (17.4.3)$$

which is continuous in x for each θ.

We further assume throughout this section that for each fixed value of x, $F(x \mid \theta)$ is strictly and continuously decreasing from 1 to 0 with θ; the respective distributions of x are said to be *stochastically ordered* when this condition holds. In the binomial example, this condition holds for both the exact and approximate formulas for $F(x \mid \theta)$, given that $x < k$. This condition is entailed by weak assumptions which are usually satisfied, namely, that each $P_g(\theta)$ increases strictly and continuously with θ, and that $x(u_1, \ldots, u_n)$ is nondecreasing in each u_g and increasing in at least one of them. The latter conditions hold in all cases described above.

When these conditions hold, the respective cdf's of scores of a given test can be represented conveniently in the manner illustrated in the schematic graphs of Figs. 17.4.1 and 17.4.2. Figure 17.4.2 is a schematic representation

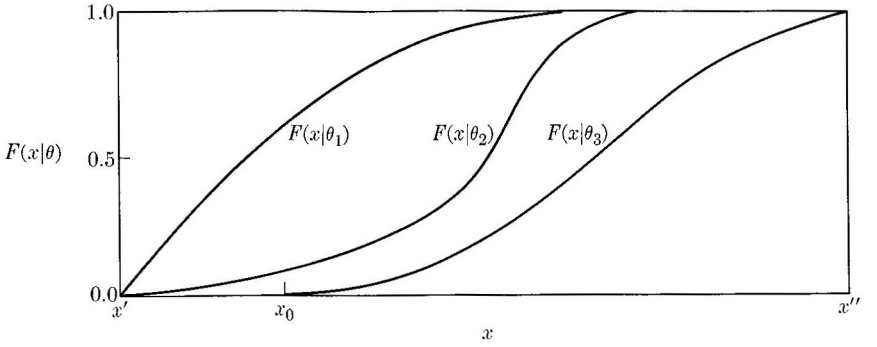

FIG. 17.4.1. Cdf's of scores x for several θ-values, $\theta_1 < \theta_2 < \theta_3$.

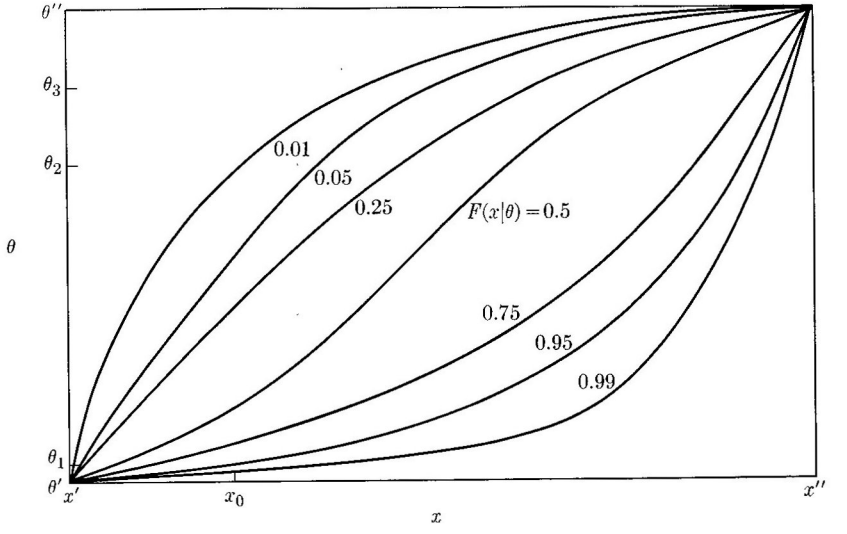

FIG. 17.4.2. Contours of constancy for cdf's $F(x, \theta)$.

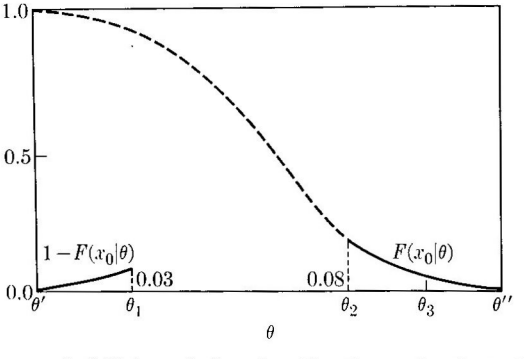

FIG. 17.4.3. Error probabilities of the classification rule that classifies high when $x > x_0$.

of the function of two arguments $F(x \mid \theta)$ that map several "contours of constant height" of the $F(x \mid \theta)$ "surface" over the (x, θ) plane. Figure 17.4.1 represents three "sections" ("slices") through this surface, made at $\theta = \theta_1, \theta_2, \theta_3$, respectively. Figure 17.4.3 represents in two forms one "section" made in the perpendicular direction at $x = x_0$; we shall explain this figure below.

The discriminating power of a test is illustrated most simply in problems of discriminating between just two levels of ability. One common rule classifies those subjects whose scores exceed some specified number x_0 as "high" and classifies others as "low". With this rule, if θ is any ability level considered definitely high, then $F(x_0 \mid \theta) = \text{Prob}\,(X \leq x_0 \mid \theta)$ is the probability of erroneous (low) classification of a subject of that ability. Since $F(x_0 \mid \theta)$ decreases as θ increases, it is natural to focus attention on the smallest θ value considered definitely high, say θ_2. The rule's maximum probability of erroneous classification of a high-ability subject is then $F(x_0 \mid \theta_2)$, as illustrated by Fig. 17.4.3. Similarly, if θ_1 is the highest ability considered definitely low, then

$$1 - F(x_0 \mid \theta_1) = \text{Prob}\,(X > x_0 \mid \theta)$$

is the rule's maximum probability of erroneous classification of a low-ability subject. At abilities between θ_1 and θ_2, neither classification is considered definitely erroneous and no error probabilities are considered.

By decreasing x_0, we can decrease $F(x_0 \mid \theta_1)$, the maximum misclassification probability for low abilities, but only at the cost of increasing $1 - F(x_0 \mid \theta_2)$, the maximum misclassification probability for high abilities. Evidently the possibility of circumventing such restrictions on the discriminating power attainable with a given test depends on basic reconsideration of the forms of test-score formulas and classification rules adopted; and these considerations might show that improvement requires the use of a different test.

We note that some of the present considerations parallel some of the interpretations given above of the discriminating power of single items in terms of item characteristic curves. The common element is the role of the rate of increase of $P_g(\theta)$ and of $1 - F(x_0 \mid \theta)$, respectively, as θ increases. So long as a test is used only to provide a classification rule based on a comparison of its scores with some fixed critical value x_0, the test is in effect equivalent to a single hypothetical test item having responses

$$u_1^* = \begin{cases} 1, & \text{corresponding to } x > x_0 \\ 0, & \text{corresponding to } x \leq x_0, \end{cases}$$

and item characteristic curve

$$P_1^*(\theta) = 1 - F(x_0 \mid \theta).$$

We can consider parameters describing the form of $F(x_0 \mid \theta)$ and $1 - F(x_0 \mid \theta)$ in rough analogy with the parameters of single items: For example, if $F(x_0 \mid \theta') = \frac{1}{2}$,

then θ' can be called the *difficulty level of the classification rule;* and

$$-\frac{\partial}{\partial\theta} F(x_0 \mid \theta)$$

evaluated at $\theta = \theta'$ can be called the *discriminating power of the classification rule.* We shall consider in detail below the ways in which such parameters of the test and other properties of classification rules depend on the parameters of the respective test items. Parameters such as a_g, b_g in logistic or normal items serve to characterize an item fully, and these parameters admit heuristically useful and relevant descriptive interpretations. However, their principal significance lies in their precise role in contributing to the information structure of a test, a notion we shall elaborate in the following sections and chapters. For a classification rule represented by a function $1 - F(x_0 \mid \theta)$, an analogous pair of parameters may be of some limited descriptive value, but in general they must fall far short of determining fully the course of $1 - F(x_0 \mid \theta)$ and the values of all error probabilities of practical interest. A summary description of the error probabilities that is more useful for many purposes is a pair of points such as those represented in Fig. 17.4.3, which indicate that at the values θ_1, θ_2 the error probabilities of respective types are 0.03 and 0.08.

A standard technique of estimation, that of confidence limits, is directly applicable when the distributions of test scores are available graphically, as in Fig. 17.4.3, or equivalently in tables of percentage points. A lower confidence limit estimator with a confidence coefficient of 95%, say, is defined as any statistic $t(\mathbf{v})$ having the property that

$$\text{Prob}\,[t(\mathbf{V}) \leq \theta \mid \theta] = 0.95 \qquad \text{for each } \theta.$$

That is, for each possible value θ of an examinee's ability, the probability is 0.95 that the estimate $t(\mathbf{v})$ derived from the response pattern of such an examinee will be a correct lower bound on his ability.

In the case at hand, where $\mathbf{v}$ is represented just by a test score x, it is easy to obtain a statistic $t(x)$ with the above property. Let x^* denote the numerical test score of an examinee. Let $\theta^*(x^*, 0.95)$ denote the number θ^* that satisfies the equation $F(x^* \mid \theta^*) = 0.95$; in Fig. 17.4.2, θ^* corresponds to x^* in the sense that (x^*, θ^*) is a point on the 0.95 contour. Then θ^* is a lower 95% confidence limit estimate of the examinee's ability θ. (The fact that $\text{Prob}\,[\theta^*(X, 0.95) \leq \theta \mid \theta] = 0.95$ is an easily derived consequence of the definition of θ^*.) Taking $\theta^* = 1.3$, for concreteness of illustration, we may record this conveniently in the notation: $\text{Conf}\,(\theta \geq 1.3) = 0.95$.

Other confidence limits are determined similarly. For example, $\theta^*(X, 0.25)$ is an upper 75% confidence limit estimator, defined implicitly by $F(x \mid \theta) = 0.25$ and having the basic property that

$$\text{Prob}\,[\theta^*(X, 0.25) > \theta \mid \theta] = 0.75 \qquad \text{for each } \theta.$$

The pair of estimators, $\theta^*(x, 0.95)$, $\theta^*(x, 0.05)$, together constitute a 90%

confidence interval estimator of θ; For each possible true value θ, they include θ between them with probability 90%. Among the various types of useful point estimators of θ, one which we may conveniently describe here is $\theta^*(x, 0.5)$. This point estimator is median-unbiased, that is, it both overestimates and underestimates θ with probability $\frac{1}{2}$.

The precision of a confidence interval estimator is represented by its confidence coefficient, together with the typical lengths of the interval estimates that it determines; or, more precisely and adequately, by error probabilities for over- or underestimation by various amounts. We shall indicate below how such precision properties of confidence intervals and confidence limits can be related in detail to the discriminatory power of a test in classification by ability levels.

17.5 The Information Structure of a Test and Transformations of Scale of Scores

When we apply a test model in conjunction with a specific test to such classification and estimation problems as the ones illustrated in the preceding section, we observe that no properties of the model play any role except the cdf's of the score of that specific test. For example, Fig. 17.4.2 might represent two different tests, each with very different numbers and types of items and item characteristic curves, but the estimation and classification methods based on scores of the respective tests would still have identical error-probability properties. This equivalence would hold even if the cdf's were different, but could be made to coincide when scores x of one test were transformed by a suitable increasing function $x^*(x)$ into scores x^* of the second test. This is true because no properties of the scale of scores x beyond simple ordering have been used here. Thus, for such standard inference methods based on an adopted test score formula, we may consider the family of distributions of scores $F(x \mid \theta)$ as representing the essential *information structure* or *canonical form* of a test, with the qualification that the scale of scores x plays only the role of simple ordering.

To illustrate this qualification, consider any given family of cdf's $F(x \mid \theta)$, any arbitrarily chosen ability θ_2, and the function defined by

$$x^* \equiv x^*(x) = F(x \mid \theta_2).$$

This is a strictly increasing function of x, and we can adopt it to define scores x^* on a new scale; the range of such scores is $0 \leq x^* \leq 1$. Let the cdf's of such scores x^* be denoted by

$$F^*(x^* \mid \theta) \equiv \text{Prob}\,[X^*(\mathbf{V}) \leq x^* \mid \theta], \qquad 0 \leq x^* \leq 1.$$

A special property of scores defined in this way is that when $\theta = \theta_2$ [that is, the ability level that has been arbitrarily chosen for the definition of $x^*(x)$], the distribution of scores X^* takes the special "uniform" form

$$F^*(x^* \mid \theta_2) \equiv x^*, \qquad 0 \leq x^* \leq 1.$$

This property characterizes the *probability integral transformation* $x^*(x)$. An illustration appears in Fig. 17.5.1, a figure that is a transformed version of Fig. 17.4.1. Since such transformations of scores are typically nonlinear, expected values and variances of the transformed scores x^* do not have any simple relations to expected values and variances of scores x on the original scale. Thus the concepts and methods presented in this section are not closely linked with any use of moments of distributions of test scores at given ability levels.

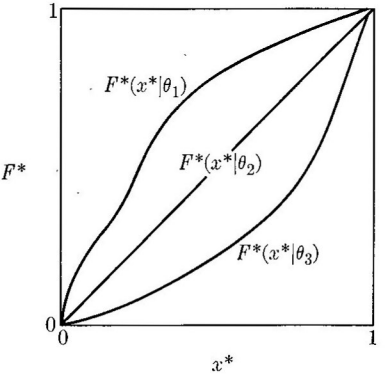

FIG. 17.5.1. Transformed version of Fig. 17.4.1.

[If the curve for θ_3 were deleted from Fig. 17.5.1 and the resulting figure were rotated, the new figure would be familiar to many students of mathematical statistics (see, for example, Lindgren, 1962, p. 236). For each α, the test of the hypothesis H_0: $\theta = \theta_2$ against the hypothesis H_1: $\theta = \theta_1$, based on rejecting H_0 just when x^* is sufficiently small, is the test that rejects just when $x^* \leq x^*_\alpha \equiv \alpha$. The power of this test is given by $1 - \beta = F^*(x^*_\alpha \mid \theta_1) \equiv F^*(\alpha \mid \theta_1)$. Thus $\beta = 1 - F^*(\alpha \mid \theta_1)$ gives the "α, β curve".]

17.6 Transformations of Scales of Ability

It is interesting to consider the preceding point concerning the scaling of *scores*, in combination with the point concerning the scaling of *abilities* illustrated at the end of Section 17.1, where a certain freedom in specification of the ability scale was discussed. The latter point can be applied here: Abilities θ can be replaced by abilities $\theta^* = \theta^*(\theta)$ on a transformed scale in such a way that the family of cdf's of scores

$$F^{**}(x^* \mid \theta^*) = \text{Prob}\,[X^*(\mathbf{V}) \leq x^* \mid \theta^*]$$

is given any chosen form compatible with the other conditions thus far assumed. For example, the transformation $\theta^*(\theta)$ defining the new scale of abilities can be chosen so that each possible score value x^* is the median of the distribution

of scores for ability level $\theta^* = \theta^*(\theta) = x^*$; that is, so that

$$F^{**}(x^* \mid \theta^*) = \tfrac{1}{2} \qquad \text{whenever} \qquad x^* = \theta^* = \theta^*(\theta),$$

as in Fig. 17.6.1, which is a transformed version of Fig. 17.4.2. To prove this, we note that the condition $F(x \mid \theta) = 0.5$ defines implicitly the function $x(\theta)$, the median of scores x for each ability θ, in terms of the given cdf's. Hence $x^*[x(\theta)]$ is the median of transformed scores x^* for each ability θ. We are now free to define a transformation of abilities by

$$\theta^*(\theta) = x^*[x(\theta)].$$

Now for each ability θ, the transformed ability $\theta^* = \theta^*(\theta)$ coincides with the median of the distribution of transformed scores.

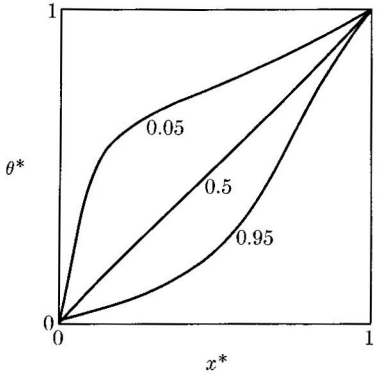

FIG. 17.6.1. Transformed version of Fig. 17.4.2.

We may mention another significant example of possible rescaling of scores and abilities. Let $x^*(x)$ be any arbitrarily chosen strictly increasing function, subject only to the mild restriction that the expected values of scores, $\mathscr{E}(X^* \mid \theta)$, exist for each θ. Let us determine a new scale of scores by the transformation $x^*(x)$. Next, we can choose a transformed scale of abilities θ^*, determined by the transformation function $\theta^*(\theta) = \mathscr{E}(X^* \mid \theta)$. From the assumption that the cdf's $F(x \mid \theta)$ are stochastically ordered, it follows that $\theta^*(\theta)$ is an increasing function, and that the cdf's $F^{**}(x^* \mid \theta^*)$ will also be stochastically ordered. Every scale of abilities θ^* that may be determined in this way satisfies the essential condition for the definition of true score presented in Chapter 2, namely, $\mathscr{E}(X^* \mid \theta^*) = \theta^*$ for each θ^* (see Chapter 24).

It is interesting to consider an analogous question: If any test model and score formula are given and are represented by a specified family of cdf's $F(x \mid \theta)$ having the two monotonicity properties assumed above, then is it always possible to keep the *given ability scaling* (which, of course, may have been obtained by an arbitrary transformation from a previous ability scaling), and also to realize simultaneously, by means of some monotone transformation $x^*(x)$ of

the score scale, the essential condition for true-score theory, namely, $\theta = \mathscr{E}(X^* \mid \theta)$ for each θ? The answer is, usually, no—the possibility depends on the detailed structure of the given cdf's $F(x \mid \theta)$. To illustrate this simply, we shall assume that x has a finite number of possible values

$$x_1 < \cdots < x_j < \cdots < x_M,$$

and consider an arbitrary sequence of different possible values of θ, namely, $\theta_1, \theta_2, \ldots \theta_i, \ldots$. Let $C_{ij} = \text{Prob}\,(X = x_j \mid \theta_i)$ for each i, j. (Here we drop the assumption of continuity of the cdf's $F(x \mid \theta)$, an assumption which is typically inexact although useful elsewhere.) If $x^*(x)$ is any monotone transformation, we may write

$$x_j^* = x^*(x_j) \qquad \text{and} \qquad x_1^* < \cdots < x_j^* < \cdots x_M^*.$$

If the transformed scores x^* are to satisfy the true score assumption

$$\theta = \mathscr{E}(X^* \mid \theta) \equiv \mathscr{E}[x^*(X) \mid \theta], \quad \text{for each } \theta,$$

then for each i we must have

$$\theta_i = \mathscr{E}[x^*(X) \mid \theta_i] \qquad \text{or} \qquad \theta_i = \sum_{j=1}^{M} C_{ij} x_j^*.$$

In general, such linear equations in M unknowns x_j^* are inconsistent, even when only $M + 1$ such equations (determined by any chosen $M + 1$ values θ_i) are considered in isolation. Thus the possibility of realizing the conditions for true score theory *for the given ability scale*, even by monotone transformation of the given score formula, is limited by and dependent on the detailed structure of the given model $F(x \mid \theta)$. This contrasts with the possibility of realizing the true score assumptions *for the given score scale*, which, as we have seen above, is always possible if a monotone transformation of the ability scale is allowed. The discussion here amplifies and formalizes the discussion in Chapter 2 of the relationship among various concepts of true score.

On the other hand, to explore the approximate applicability of classical true-score theory to a given model when the given ability scaling is to be retained, we can first choose successively values θ_i that seem to represent the range of abilities of interest effectively and can then consider the sequence of equations

$$\theta_i = \sum_{j=1}^{M} C_{ij} x_j^*,$$

continuing so long as the equations are consistent and allow ordered solutions x_j^*. If any set of such equations does not determine unique, ordered solutions, we may supplement it by adding arbitrary and possibly convenient independent

linear restrictions on the x_j^*, possibly including specification of convenient values for

$$x_1^*, \quad x_M^*, \quad \frac{x_1^* + x_M^*}{2}, \quad \text{or} \quad \frac{1}{M}\sum_{j=1}^{M} x_j^*,$$

or some combination of these, until we have obtained M linearly independent equations.

Whenever $F(x \mid \theta)$ is a normal cdf for each θ, we may take $\theta^*(\theta) = \mathscr{E}(X \mid \theta)$, which is both the mean and the median of X, for each θ. When $F(x \mid \theta)$ is at least approximately a normal cdf, then $\theta^*(\theta) = \mathscr{E}(X \mid \theta)$ is usually approximately the median (as well as exactly the mean) of X.

Of course, weak true-score theory is characterized by its use of no restrictive assumptions on the forms of the cdf's $F(x \mid \theta)$ of scores other than low-order moments of scores. The preceding considerations illustrate some of the many connections and differences to be found between weak and strong true-score theories.

17.7 Calculations of Distributions of Test Scores

Applications of the inference methods illustrated above require adequate numerical determinations of the distributions of test scores at respective ability levels. In most practical work with cognitive tests, response patterns are represented only by test scores having the particular form

$$x = x(\mathbf{v}) = \sum_{g=1}^{n} w_g u_g \tag{17.7.1}$$

of weighted sums of item responses, where the w_g are specified numerical weights. Most commonly, the weights are specified as equal, either as $w_g \equiv 1$, where calculation of x then gives the number of correct responses, or as $w_g \equiv 1/n$, where calculation of x then gives the proportion of correct responses. In the following chapters, we shall see that in important cases a suitably chosen linear (or weighted-sum) score formula can be used to provide estimators with optimal or nearly optimal precision and classification rules of good discriminating power. In this section, we shall present some useful theoretical and computational methods for calculating distributions of test scores of this form and illustrate these by numerical examples of the applications illustrated above.

The principal result that we shall present here is the normal approximation to the cdf $F(x \mid \theta)$ for score formulas x of any weighted sum form $x = \sum_{g=1}^{n} w_g u_g$, where the w_g are given constants. The theoretical basis for such normal approximations in the general case consists of the central limit theorems available for sums of nonidentical independent random variables (see, for example, Lindgren, 1962, p. 147, and Loève, 1955, p. 288). The resulting approximation formulas for $F(x \mid \theta)$ depend on the given test model only through the mean

and variance of X for each θ:

$$\mathscr{E}(X \mid \theta) = \sum_{g=1}^{n} \mathscr{E}(w_g U_g \mid \theta) \equiv \sum_{g=1}^{n} w_g P_g(\theta), \tag{17.7.2}$$

$$\sigma^2(X \mid \theta) = \sum_{g=1}^{n} \sigma^2(w_g U_g \mid \theta) \equiv \sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta). \tag{17.7.3}$$

Then the approximation formula is

$$F(x \mid \theta) \doteq \Phi\{[x - \mathscr{E}(X \mid \theta)]/\sigma(X \mid \theta)\}. \tag{17.7.4}$$

In connection with various specific test models and problems of application below, the preceding general formulas for the moments of scores will be specialized and substituted in the last relation.

For test models with nonequivalent items, and for composite scores with unequal weights, we require here a form of the central limit theorem that allows nonidentically distributed terms $w_g U_g$. On the other hand, for many practical purposes we may conveniently interpret the hypothetical concept of increase without limit of the number n of nonequivalent test items as the case of a test model with $G_n = ng$ items specified as follows: The first n items may have any specified ICCs; each successive set may consist of n items equivalent, respectively, to those of the first set; and G may increase without limit. The simplest case of the central limit theorem, that of identically distributed terms, applies here, since each set of n items can formally be considered to contribute a single term

$$Z_r = \sum_{g=1}^{n} w_g U_{nr+g}, \quad r = 0, 1, 2, \ldots, \qquad \text{to} \qquad X = \sum_{r=0}^{G} Z_r,$$

provided that $w_{nr+g} = w_g$ for $r = 1, 2, \ldots$.

Examples: Moments and quantiles of test scores for items of various types.

Moments of item responses

$$\mathscr{E}(U_g \mid \theta) = P_g(\theta), \qquad \sigma^2(U_g \mid \theta) = P_g(\theta) Q_g(\theta),$$

1. Normal ogive

$$\mathscr{E}(U_g \mid \theta) = \Phi[L_g(\theta)], \qquad \sigma^2(U_g \mid \theta) = \Phi[L_g(\theta)]\Phi[-L_g(\theta)],$$

where $L_g(\theta) = a_g\theta - b_g$.

2. Logistic

$$\mathscr{E}(U_g \mid \theta) = \Psi[DL_g(\theta)], \qquad \sigma^2(U_g \mid \theta) = \psi[DL_g(\theta)],$$

where

$$\psi(t) = \frac{\partial}{\partial t}\Psi(t) \equiv \frac{e^t}{(1 + e^t)^2}.$$

3. Three-parameter logistic

$$\mathscr{E}(U_g \mid \theta) = c_g + (1 - c_g)\Psi[DL_g(\theta)] = \Psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)],$$
$$\sigma^2(U_g \mid \theta) = (1 - c_g)\psi[DL_g(\theta)] + c_g(1 - c_g)\Psi[-DL_g(\theta)]^2.$$

Moments of terms in locally best composite scores (developed below in Section 19.3)

$$w_g(\theta) = P'_g(\theta)/P_g(\theta)Q_g(\theta),$$
$$\mathscr{E}[w_g(\theta)U_g \mid \theta] = [P'_g(\theta)/P_g(\theta)Q_g(\theta)]P_g(\theta) = P'_g(\theta)/Q_g(\theta),$$
$$\sigma^2[w_g(\theta)U_g \mid \theta] = w_g(\theta)^2\sigma^2(U_g \mid \theta) = [P'_g(\theta)^2/P_g(\theta)^2Q_g(\theta)^2]P_g(\theta)Q_g(\theta)$$
$$= P'_g(\theta)^2/P_g(\theta)Q_g(\theta).$$

1. Normal ogive

$$w_g(\theta) = a_g\varphi[L_g(\theta)]/\Phi[L_g(\theta)]\Phi[-L_g(\theta)],$$
$$\mathscr{E}[w_g(\theta)U_g \mid \theta] = \varphi[L_g(\theta)]/\Phi[-L_g(\theta)],$$
$$\sigma^2[w_g(\theta)U_g \mid \theta] = a_g^2\varphi[L_g(\theta)]^2/\Phi[L_g(\theta)]\Phi[-L_g(\theta)].$$

2. Logistic

$$w_g(\theta) = Da_g \quad \text{(uniformly best weights)},$$
$$\mathscr{E}(w_gU_g \mid \theta) = Da_g\psi[DL_g(\theta)], \qquad \sigma^2(w_gU_g \mid \theta) = D^2a_g^2\psi[DL_g(\theta)].$$

3. Three-parameter logistic

$$w_g(\theta) = Da_g\Psi[DL_g(\theta) - \log c_g], \qquad \mathscr{E}[w_g(\theta)U_g \mid \theta] = Da_g\Psi[DL_g(\theta)],$$
$$\text{Var}\,[w_g(\theta)U_g \mid \theta] = (1 - c_g)D^2a_g^2\psi[DL_g(\theta) - \log c_g]\psi[DL_g(\theta)].$$

Moments of composite scores. Dividing each weight w_g in a scoring formula by the same positive constant (for example, the sum of the weights) does not change the ratio between respective weights, which is the essential feature of the scoring formula. Therefore we may express any composite score formula in the form

$$x = \frac{\sum_{g=1}^{n} w_g u_g}{\sum_{g=1}^{n} w_g}.$$

For example, the weights w_g may be .

1) equal weights, for instance, $w_g = 1$ for $g = 1, \ldots, k$;
2) best weights (developed below in Section 18.4)

$$w_g(\theta_1, \theta_2) = \log \frac{P_g(\theta_2)Q_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_2)};$$

or

3) locally best weights $w_g(\theta) = P_g'(\theta)/P_g(\theta)Q_g(\theta)$ as developed in Section 19.3.

Thus we may write the moments of any composite score x as, say,

$$\mathscr{E}(X \mid \theta) = \frac{\sum_{g=1}^{n} w_g P_g(\theta)}{\sum_{g=1}^{n} w_g \equiv \mu(\theta)} \quad \text{and} \quad \sigma^2(X \mid \theta) = \frac{\sum_{g=1}^{n} w_g^2 P_g(\theta)Q_g(\theta)}{\left(\sum_{g=1}^{n} w_g\right)^2} \equiv \sigma^2(\theta).$$

Quantiles of composite scores under the normal approximation: a measure of information. Using our previous assumption that a given composite score x has cdf's $F(x \mid \theta)$ that are continuously strictly increasing in x and decreasing in θ, we may implicitly define the $(1 - \alpha)$-quantile of X, which we denote by

$$x^*(1 - \alpha, \theta),$$

as the solution x of

$$F(x \mid \theta) = 1 - \alpha. \tag{17.7.5}$$

If we assume in particular a normal form for the cdf's $F(x \mid \theta)$, we have

$$x^*(1 - \alpha, \theta)$$

defined as the solution x of

$$F(x \mid \theta) \equiv \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = 1 - \alpha. \tag{17.7.6}$$

Taking Φ^{-1} of both sides and solving for x, we then have

$$x^*(1 - \alpha, \theta) = \mu(\theta) + \Phi^{-1}(1 - \alpha)\sigma(\theta). \tag{17.7.7}$$

[The quantity $\Phi^{-1}(1 - \alpha)$ is a normal deviate cutting off a normal-curve left-tail area of $1 - \alpha$.]

The composite score will actually approach a normal form with increasing n, under the slight restriction that the values $w_g^2 P_g(\theta)Q_g(\theta)$ are uniformly bounded away from zero for the given θ-value considered. (This follows from the central limit theorem for the case of nonidentically distributed terms; see, for example, Loève, 1955, p. 310.) Under mild additional conditions (which will often be satisfied, and which can be checked with reference to specific applications), formula (17.7.7) can be approximated adequately closely, over any interval of θ-values centered at any given value θ' and appreciably wide, by a linear function of θ. This function of θ may be written

$$x^*(1 - \alpha, \theta) = A + B(\theta - \theta'), \tag{17.7.8}$$

where

$$A = \mu(\theta') + \Phi^{-1}(1 - \alpha)\sigma(\theta') \quad \text{and} \quad B = \mu'(\theta'), \quad \mu' = \frac{\partial}{\partial\theta}\mu(\theta).$$

[This function represents the Taylor series approximation to (17.7.7),

$$x^*(1 - \alpha, \theta) \doteq \mu(\theta') + \Phi^{-1}(1 - \alpha)\sigma(\theta') + \mu'(\theta')(\theta - \theta') + \Phi^{-1}(1 - \alpha)\sigma'(\theta')(\theta - \theta'),$$

further simplified by deleting the last term. This term may be deleted because

$$\sigma'(\theta') = \frac{\partial}{\partial\theta}\sigma(\theta)|_{\theta=\theta'}$$

tends to be negligible in comparison with $\mu'(\theta')$.] By solving (17.7.8) for θ, we obtain the corresponding linear approximation

$$\theta^*(x, 1 - \alpha) = \theta' + [x - \mu(\theta') - \Phi^{-1}(1 - \alpha)\sigma(\theta')]/\mu'(\theta'). \tag{17.7.9}$$

Now the latter formula represents (approximately) the lower $(1 - \alpha)$-level confidence limit estimate of the ability θ of an individual with score x, as discussed in Section 17.4 above. One natural and convenient indication of the value of a given test and scoring formula is the width of the resulting confidence interval estimates of ability. The width of the approximate $(1 - 2\alpha)$-level confidence interval indicated by the approximation (17.7.9) for any given $\alpha < \frac{1}{2}$ is just $\theta^*(\alpha, x) - \theta^*(1 - \alpha, x)$ as determined from (17.7.9):

$$[\Phi^{-1}(1 - \alpha) - \Phi^{-1}(\alpha)]\frac{\sigma(\theta')}{\mu'(\theta')}.$$

We see that this width is proportional to $\sigma(\theta')/\mu'(\theta')$, a constant independent of α, x, and θ under the assumed approximation. For θ near θ', therefore, this constant serves as an index of precision of interval estimation based on the given test and scoring formula. As it turns out, the same constant also characterizes the effectiveness of the test and scoring formula for a wide variety of other purposes. Hence we shall use the term *information* to designate the related quantity

$$I(\theta', x) = \mu'(\theta')^2/\sigma^2(\theta'). \tag{17.7.10}$$

More precisely, we shall refer to $I(\theta', x)$ as the *information* provided by the given test and composite scoring formula in the neighborhood of θ'. The function $I(\theta, x)$ is called *the information function of the scoring formula* x. It should be noted that the symbol x appears here not as a variable argument of $I(\theta, x)$, but as an abbreviation for "the probability distributions $F(x, \theta)$ of the scoring formula x", in terms of which $I(\theta, x)$ is defined. The definition is of course made with reference to some specified mental test, in terms of which the scoring formula is defined; thus, for a given scoring formula, $I(\theta, x)$ is a function of θ only.

There are two additional reasons for using the term "information" for this quantity:

1. Let us consider the error probability functions of classification rules based on x. At θ', the slope of each of these functions has a given value $\alpha = \alpha(\theta')$. With increasing n, these values tend to be proportional to

$$\sqrt{I(\theta', x)} \equiv \mu'(\theta')/\sigma(\theta').$$

Proof. Writing $F(x \mid \theta) = \Phi\{[x - \mu(\theta)]/\sigma(\theta)\} = \Phi(t)$, where $t = [x - \mu(\theta)]/\sigma(\theta)$, we see that the slope of the error probability function $1 - F(x \mid \theta)$ is $-\partial F(x \mid \theta)/\partial\theta$. By the chain rule, this can be written as $-[d\Phi(t)/dt]\,(\partial t/\partial\theta)$, or $-\varphi(t)\,(\partial t/\partial\theta)$. Hence

$$\frac{\partial t}{\partial\theta} = -\frac{\{\sigma(\theta)\mu'(\theta) + [x - \mu(\theta)]\sigma'(\theta)\}}{\sigma^2(\theta)} = -\frac{\mu'(\theta)}{\sigma(\theta)} - t\frac{\sigma'(\theta)}{\sigma(\theta)}$$
$$\doteq -\sqrt{I(\theta, x)},$$

since, as we noted in the derivation of (17.7.9), $\sigma'(\theta)$ is small compared with $\mu'(\theta)$.

2. With increasing n, when θ' is the true value, the point estimator $\theta^*(x, 0.5)$ tends to be normally distributed, with mean θ' and variance $1/I(\theta', x)$.

Thus $I(\theta, x)$ plays the role of an index of precision of estimation. If we are dealing with nonlinear scoring functions $x = x(\mathbf{v})$, then we cannot apply the central limit theorem in the direct way indicated in connection with (17.7.4) above. Nevertheless, for an important and wide class of nonlinear scoring functions and estimators, we can show that there is an approach to a limiting normal distribution with increasing n. The definition $I(\theta, x)$ or $I(\theta, \theta^*) = \mu'(\theta)^2/\sigma^2(\theta)$ is extended to such cases of nonlinear x or θ^* by taking $\mu(\theta)$ and $\sigma^2(\theta)$ to represent the *asymptotic moments* of x and θ^*. These asymptotic moments are moments of the limiting normal distributions, which are in theory, and in relevant examples, distinct from the limits of exact moments of x or θ^*.

In particular, in Section 20.3, we shall consider the maximum likelihood estimator $\hat{\theta}$ and its information function $I(\theta, \hat{\theta})$ in some detail. As in the preceding special case, we shall see that the role of an index of precision of estimation is played quite frequently by the information function $I(\theta, x)$, for a given scoring formula, and $I(\theta, \theta^*)$, for a given test and estimator.*

---

* For derivations and discussions of these properties, see Cramér (1946, pp. 498–506) or Birnbaum (1961a, pp. 122–127). In such discussions of asymptotic distributions in connection with maximum likelihood, the results (1) and (2) above are obtained by replacing the "score" $S(x, \theta) = (\partial/\partial\theta) \log f(x \mid \theta)$ by $[x - \mathscr{E}(X \mid \theta)]/\sigma(X \mid \theta)$.

These and other uses and interpretations of the information functions $I(\theta, x)$ of various test models and composite score formulas will appear below, particularly in Chapter 20, where self-contained discussions of some aspects of information functions are given.

17.8 Quantal Response Models in General

The test models introduced in this chapter have analogues in other technical and scientific areas. Models of the general form Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ have been called *quantal response models.* The normal ogive model (including the three-parameter case described above) has been used extensively in biological assay work. (See, for example, Finney, 1944 and 1952. In the second reference, comparisons between biological and psychometric applications are given.) The use of the logistic model as an alternative to the normal in bioassay work has also been developed extensively (see Berkson, 1953 and 1957). For another type of biological assay, the *dilution series* model with $P_g(\theta) = 1 - e^{-a_g\theta}$ has been used (Fisher, 1922, pp. 363–366, and Cochran, 1950). Applications of such models have also been made in industrial gauging (Stevens, 1948) and genetics (for example, Rao, 1965, pp. 302–309, and Kempthorne, 1957, p. 181, and references therein). An appreciable part of the discussion in the next chapters has general relevance to quantal response models.

17.9 Estimation of Item Parameters

Two maximum likelihood methods have been given for estimating the item parameters in the normal ogive test model, by Tucker (1951) and by Lord (1953). These are discussed by Torgersen (1958, pp. 388–391), where they are related to other mathematical problems that arise in scaling. In the following paragraphs (1) and (2), we present two adaptations of these methods to the case of the logistic model. [For the restricted case of the logistic model described in Section 17.2, in which only the item difficulty parameters b_g are unknown, Rasch (1960) has given advantageous estimation methods. Many details of the derivation and calculation of estimates presented in the next paragraphs have forms similar to those of the more restricted estimation problem discussed in more detail in Section 20.3, which deals with maximum likelihood estimates of ability.]

The likelihood function of the responses observed when an n-item test is administered to a group of N examinees of abilities $\theta_1, \theta_2, \ldots, \theta_N$ is

$$L = \prod_{c=1}^{N} \prod_{g=1}^{n} \{1 - \Psi[Da_g(\theta_c - b_g)]\} \exp[Da_g(\theta_c - b_g)u_{gc}]. \qquad (17.9.1)$$

Let

$$x_c = \sum_{g=1}^{n} u_{gc}$$

denote the raw score of examinee c. Then

$$\mathscr{E}(X_c \mid \theta_c) = \sum_{g=1}^{n} \Psi[Da_g(\theta_c - b_g)]$$

is an increasing function of θ_c, provided that all a_g are positive. For two examinees of abilities θ_c and $\theta_{c'} > \theta_c$, we have Prob $\{X_{c'} > X_c\} \to 1$ as n increases, provided that the a_g are bounded away from zero and the b_g are bounded. That is, there is a tendency for ability order to be reflected correctly in the ordering of raw scores, as the number of items increases.

1. If we assume that the examinees are a random sample from a population in which the ability θ has a standard normal (or logistic) distribution, then, as N increases, the distribution of θ_c values over examinees converges (with probability one) to the standard normal (or logistic) distribution. Correspondingly the ability $\theta_{[PN]}$, which exceeds just a given proportion P of the abilities θ_c in a sample of n examinees, converges (with probability one), as n increases, to $\theta_P = \Phi^{-1}(P)$ [or to $\Psi^{-1}(P)/D$]. This second limit is the ability that exceeds just the proportion P of abilities in the population. Let P_c denote the proportion of raw scores in the sample that are less than x_c, and let

$$\theta(x_c) = \Psi^{-1}(P_c)/D. \tag{17.9.2}$$

Then it follows, under the conditions on item parameters mentioned in the preceding paragraph, that $\theta(x_c) \to \theta_c$ (with probability one) as both n and N increase. Thus, in practice, with N and n finite, we may regard $\theta(x_c)$ as an estimate of θ_c. In the next paragraphs, we treat the θ_c as known, with the understanding that in applications they shall be replaced by their numerical estimates $\theta(x_c)$.

The likelihood function L now has as unknown arguments just the $2n$ item parameters a_g and b_g. The maximum likelihood equations

$$\frac{\partial \log L}{\partial a_g} = 0, \qquad \frac{\partial \log L}{\partial b_g} = 0,$$

are easily simplified to

$$\frac{1}{N}\sum_{c=1}^{N} \theta_c \Psi[Da_g(\theta_c - b_g)] = t_g, \qquad g = 1, \ldots, n, \tag{17.9.3}$$

$$\frac{1}{N}\sum_{c=1}^{N} \Psi[Da_g(\theta_c - b_g)] = s_g, \qquad g = 1, \ldots, n, \tag{17.9.4}$$

where

$$s_g = \frac{1}{N}\sum_{c=1}^{N} u_{gc} \qquad \text{and} \qquad t_g = \frac{1}{N}\sum_{c=1}^{N} \theta_c u_{gc}.$$

For each g, the pair of equations (17.9.3) and (17.9.4) in a_g and b_g can be solved for the maximum likelihood estimates $\hat{a}_g$ and $\hat{b}_g$ by numerical iteration with the aid of Berkson's (1957) tables of Ψ.

After each cycle, or after several cycles, of calculation of the successive approximation values

$$[a_g^{(1)}, b_g^{(1)}], \ldots, [a_g^{(r)}, b_g^{(r)}], \qquad g = 1, \ldots, n,$$

the first trial values

$$\theta_c^{(1)} = \theta(x_c) \tag{17.9.5}$$

given by (17.9.2) may be replaced by the successive approximations $\hat{\theta}_c^{(r)}$, for $c = 1, \ldots, n$, where $\hat{\theta}_c^{(r)}$ is a formal solution of the equation for estimation of θ_c when all item parameters are assumed known. This formal solution and its conditions are discussed in detail in Section 20.3 below and used in the next paragraph.

2. Dropping now the assumption made in (1) of a known prior distribution of abilities, we may obtain from L the maximum likelihood estimates $\hat{\theta}_c$ of the examinees' abilities θ_c, along with the estimates $\hat{a}_g$ and $\hat{b}_g$ of item parameters. Even in this case it is convenient to begin an iterative procedure for computing all $\hat{\theta}_c$, $\hat{a}_g$, and $\hat{b}_g$ with first-cycle values $\theta_c^{(1)} = \theta(x_c)$ defined as in (17.9.2). Then second-cycle values $\theta_c^{(2)}$ can be obtained from the maximum likelihood equation (see Section 20.3)

$$\partial \log L / \partial \theta_c = 0,$$

or

$$\sum_{g=1}^{n} a_g \Psi[D a_g(\theta_c - b_g)] = \sum_{g=1}^{n} a_g u_{gc}, \tag{17.9.6}$$

with a_g and b_g replaced by $a_g^{(1)}$ and $b_g^{(1)}$. Then $\hat{\theta}^{(1)}$ can be replaced by $\hat{\theta}^{(2)}$ in (17.9.3) and (17.9.4), and the second-cycle values $a_g^{(2)}$ and $b_g^{(2)}$ can be obtained as solutions of those equations. Further cycles could run through (17.9.6), (17.9.3), and (17.9.4) in several possible patterns of iteration.

Lord (1967) has successfully applied a procedure similar to that just outlined to various sets of data, using a computer program written by Diana Lees. In one application, the a_g, b_g, and θ_c values were simultaneously estimated for 3000 examinees and 90 items (a total of 270,000 item responses). Bock (1967) has reported successful estimation of a_g and b_g values by a method based on the assumption that θ_c is normally distributed in the population of examinees. Substantial variation in a_g values was found in both of these applications.

17.10 Validity of Test Models

Some aspects of questions of validity and adequacy of fit of specific test models were discussed in Chapter 16. For the logistic model, the estimation methods indicated above may be useful as part of an empirical test of fit. Where specific

techniques of testing fit are concerned, the reader should be aware that some established approaches to testing goodness of fit have come to be considered unsound and potentially misleading by a number of statisticians and scientific workers. An alternative perspective on testing adequacy of models is one based primarily on rather direct, often graphical, comparisons of data with significant aspects of models. Here a crucial role is played by relatively unformalized judgments that involve both the subject-matter context and statistical considerations. Bush (1963) has described and illustrated one such perspective on testing models.

The bearing of some of these questions on statistical efficiency of estimation of ability will be discussed in Section 19.1.

References and Selected Readings

BERKSON, J., A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, 1953, **48,** 565–599.

BERKSON, J., Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, 1957, **13,** 28–34.

BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.

BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29,** 1284 (abstract). (c)

BIRNBAUM, A., Statistical theory of tests of a mental ability. *Op. cit.*, 1285 (abstract). (d)

BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics*, 1961, **32,** 112–135. (a)

BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)

BIRNBAUM, A., Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin 67–12.* Princeton, N.J.: Educational Testing Service, 1967.

BOCK, R. D., Fitting a response model for n dichotomous items. Paper read at the Psychometric Society Meeting, Madison, Wisconsin, March 1967.

BUSH, R. B., *Handbook of mathematical psychology*, Vol. 1, Chapter 8: Estimation and evaluation. New York: Wiley, 1963.

COCHRAN, W. G., Estimation of bacterial densities by means of the most probable number. *Biometrics*, 1950, **6,** 105–116.

CRAMÉR, H., *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press, 1946.

FINNEY, D. J., The application of probit analysis to the results of mental tests. *Psychometrika*, 1944, **9,** 31–39.

FINNEY, D. J., *Probit analysis.* London: Cambridge University Press, 1952.

FISHER, R. A., On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London* (A), 1922, **222,** 309–368. (Reprinted in R. A. Fisher, *Contributions to mathematical statistics.* New York: Wiley, 1950.)

GUTTMAN, L., Chapters 2, 3, 6, 8, 9 in S. A. Stouffer *et al., Measurement and prediction.* Princeton, N.J.: Princeton University Press, 1950.

HALEY, D. C., Estimation of the dosage mortality relationship when the dose is subject to error. *Technical Report No. 15,* August 29, 1952. Stanford, Calif.: Contract No. ONR-25140, Applied Mathematics and Statistics Laboratory, Stanford University.

KEMPTHORNE, O., *An introduction to genetic statistics.* New York: Wiley, 1957.

LAZARSFELD, P., Latent structure analysis. In S. Koch (Ed.), *Psychology: a study of a science,* Vol. 3. New York: McGraw-Hill, 1959.

LINDGREN, B. W., *Statistical theory.* New York: Macmillan, 1960, 1962.

LOÈVE, M., *Probability theory.* New York: Van Nostrand, 1955.

LORD, F. M., A theory of test scores. *Psychometric Monograph, No. 7.* Chicago: University of Chicago Press, 1952. (a)

LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17,** 181–194. (b)

LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18,** 57–76.

LORD, F. M., An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Research Bulletin 67–34.* Princeton, N.J.: Educational Testing Service, 1967.

MCNEMAR, Q., *Psychological statistics.* New York: Wiley, 1962.

RAO, C. R., *Linear statistical inference and its applications.* New York: Wiley, 1965.

RASCH, G., *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut), 1960.

STEVENS, W. L., Control by gauging. *Journal of the Royal Statistical Society* (B), 1948, **10,** 54–108.

TORGERSON, W. S., *Theory and methods of scaling.* New York: Wiley, 1958.

TUCKER, L. R, Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11,** 1–14.

TUCKER, L. R, Academic ability test. *Research Memorandum 51–17.* Princeton, N.J.: Educational Testing Service, 1951.

CHAPTER 18

TEST SCORES, SUFFICIENT STATISTICS, AND THE INFORMATION STRUCTURES OF TESTS

18.1 Sufficient Statistics: Definition and Interpretation

In applications, the test models discussed in the preceding chapter principally serve as frames of reference for analyzing and interpreting the information that an examinee's response pattern $\mathbf{v}$ provides about his ability level θ. In this chapter and in Chapters 19 and 20, we shall discuss some general concepts that give exact meaning to several common-sense notions of information. These concepts in turn will guide our consideration of specific applied techniques.

We introduce these concepts on the assumption that a given test model can be known to be valid. In practice we never know whether a given model is precisely valid, and it seems doubtful on general grounds that any model that has one of the special forms considered in the preceding chapter would be precisely valid in any application. Hence it is important to consider the typical conditions under which these concepts and their implications are applied. Under such conditions there is typically only incomplete statistical and theoretical evidence, and this evidence can therefore support only the approximate validity of models. Such considerations have not yet been developed systematically in test theory. Indeed, even for the standard statistical problems of combination and adjustment of observations, some of these considerations have been taken up systematically only in the present decade; notably, development of *efficiency-robust* methods of estimation. Thus our presentation will largely be concerned with what may be called classical statistical concepts and techniques. We may expect these to have some permanent and general value for theory and practice, notwithstanding the indicated need for further basic developments.

The greatest possible simplification of data without loss of information is an important goal in many areas of applied statistics, and particularly in the theory and practice of test scoring. A given set of statistical observations, such as response patterns, may be considered simplified if it is represented by suitable statistics that are easier to use and interpret than the original data. Technically a *statistic* is defined as any function of a response pattern, possibly a vector-valued function. Any score formula $t = t(\mathbf{v})$ is an example of a (real-valued) statistic.

Any score formula $t(\mathbf{v})$ seems to provide real simplification, since its range is one-dimensional while the range of $\mathbf{v}$ is n-dimensional. However, one hopes that we may determine score formulas that preserve all the information in response patterns, and also order the response patterns appropriately, that is, according to apparently increasing ability. We shall consider the latter desideratum in detail in Chapter 19. In this chapter we consider the former, simplification, in a general and basic way, in terms of statistics that are not necessarily score formulas or real-valued statistics.

The concept of a sufficient statistic, as we shall develop it here, is a precise version of familiar notions of the information in a message or in a set of data. In a given context, it may be that many detailed aspects of a message or a set of data are irrelevant: If they were changed or ignored, there would be no change or loss of information. The concept of a *sufficient statistic* formalizes the notion of abbreviations and deletions that entail no loss of information. The concept of a *minimal* sufficient statistic formalizes the notion of the greatest simplification possible without loss of information.

We shall see that the simple score formula $\bar{y}$ (proportion correct) is a minimal sufficient statistic in exactly two cases: (1) test models that have equivalent items and a common ICC of any form, and (2) the general logistic test model.

Taking any statistic $\mathbf{s} = \mathbf{s}(\mathbf{v})$, we can rewrite the mathematical model of the test:

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \text{Prob}\,[\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) \mid \theta]\,\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}), \theta]. \tag{18.1.1}$$

This is a special case of the general probability formula

$$\text{Prob}\,(A \text{ and } B \mid \theta) = \text{Prob}\,(A \mid \theta)\,\text{Prob}\,(B \mid A, \theta),$$

which is obtained by considering A to be the event $\mathbf{s}(\mathbf{V}) = \mathbf{s}(\mathbf{v})$ and by considering B to be the event $\mathbf{V} = \mathbf{v}$, and by noting that in this case both A and B are true if and only if B is true, and therefore that

$$\text{Prob}\,(A \text{ and } B \mid \theta) = \text{Prob}\,(B \mid \theta).$$

For some, but not all, statistics $\mathbf{s}(\mathbf{v})$, the conditional probability $\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{s}(\mathbf{V}) = \mathbf{s}(\mathbf{v}), \theta]$ is found to be independent of θ for each possible $\mathbf{v}$. With such a statistic, we can denote the conditional pdf simply by

$$\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{V})].$$

Equation (18.1.1) then assumes the form

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \text{Prob}\,[\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) \mid \theta]\,\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})]. \tag{18.1.2}$$

Any statistic $\mathbf{S}(\mathbf{V})$ for which this holds is called a *sufficient statistic.*

To illustrate and justify the interpretation and application of sufficient statistics, we may think of the observation of a subject's response pattern $\mathbf{V} = \mathbf{v}$ as being carried out in two stages. First an observation $\mathbf{s}$ is taken of just the value of a given statistic: $\mathbf{S} \equiv \mathbf{S}(\mathbf{V})$. Then an observation is taken from among those response patterns $\mathbf{v}$ compatible with the observed value $\mathbf{s}(\mathbf{v}) = \mathbf{s}$. Mathematical models of these two hypothetical stages in the observation of $\mathbf{V} = \mathbf{v}$ are in fact given by the respective factors of Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ in its rewritten form above (Eq. 18.1.2). We may interpret the mathematical condition of sufficiency of $\mathbf{S}(\mathbf{V})$, that the second factor be independent of θ, to mean that just the value $\mathbf{s}(\mathbf{v}) = \mathbf{s}$ (but not $\mathbf{v}$ itself) is reported as it is determined from the response pattern $\mathbf{v}$ of a subject; for example,

$$\mathbf{v}' = (\mathbf{u}_1, \mathbf{u}_2) \qquad \text{and} \qquad \mathbf{s}(\mathbf{v}) \equiv u_1 + u_2 = 1.$$

As an alternative to the further precise determination of an examinee's response pattern, consider spinning a roulette wheel with outcome labels $\mathbf{v}$ having respective known probabilities Prob $[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})]$ independent of θ. For example,

$$\text{Prob}\,[\mathbf{V}' = (1, 0) \mid U_1 + U_2 = 1] = \tfrac{1}{3};$$

hence

$$\text{Prob}\,[\mathbf{V}' = (0, 1) \mid U_1 + U_2 = 1] = \tfrac{2}{3}.$$

The roulette wheel and its observed outcome $\mathbf{v}$ seem obviously irrelevant to inferences about θ, since θ does not now play a part in determining $\mathbf{V} = \mathbf{v}$. However, this merely illustrates vividly that the further determination of the subject's response pattern $\mathbf{v}$, after determination of the value of $\mathbf{s}(\mathbf{v})$, is similarly irrelevant. Thus, when $\mathbf{S}(\mathbf{V})$ is a sufficient statistic, the absence of mathematical dependence of

$$\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v})]$$

on θ characterizes the *irrelevance*, for statistical inferences concerning θ, of an observation on $\mathbf{V}$, given the observed value of $\mathbf{s}(\mathbf{V})$. We may say that for given $\mathbf{S}(\mathbf{V})$, $\mathbf{V}$ is not an indicant of θ.

This concept might be illustrated in further concrete detail by showing that any estimator or classification rule based on $\mathbf{v}$ can be matched exactly in all its error-probability properties (in fact, in all probability properties) by an estimator or classification rule based on the statistic $\mathbf{S} = \mathbf{S}(\mathbf{V})$ and not otherwise dependent on $\mathbf{V}$.

A statistic $\mathbf{s}(\mathbf{v})$ is called *minimal sufficient* if it is sufficient, and if it is a single-valued function

$$\mathbf{S}(\mathbf{V}) = \mathbf{t}[\mathbf{Z}(\mathbf{V})]$$

of every other sufficient statistic. Thus a minimal-sufficient statistic assumes common values on the largest possible sets of points $\mathbf{v}$ compatible with sufficiency. Such a statistic always exists and is unique, except for one-to-one transformations. (The models considered here, which have discrete sample

spaces, always satisfy the assumptions that guarantee existence; the reader should see Lindgren, 1962, pp. 196–200, for example.) Hence we may refer to *the* minimal sufficient statistic.

18.2 Conditions for Sufficiency of a Statistic

We observe that if $\mathbf{S}(\mathbf{V})$ is any sufficient statistic, then the factored form of the probability function of $\mathbf{V}$,

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \text{Prob}\,[\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) \mid \theta]\,\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{s}(\mathbf{V}) = \mathbf{s}(\mathbf{v})],$$

contains only one factor that is dependent on θ, and this factor is not dependent on $\mathbf{v}$ except through the value of $\mathbf{s}(\mathbf{v})$. This observation provides us with a convenient criterion for the sufficiency of a statistic. Conversely, let us suppose that $\mathbf{s}(\mathbf{v})$ is any statistic satisfying the condition that the probability function of $\mathbf{V}$ can be written in a factored form in which the factor dependent on θ is not dependent on $\mathbf{v}$ except through the value of $\mathbf{s}(\mathbf{v})$:

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = g[\mathbf{s}(\mathbf{v}), \theta]h(\mathbf{v}). \tag{18.2.1}$$

Then from (18.1.1) we have

$$\begin{aligned}
\text{Prob}\,[\mathbf{V} = \mathbf{v} \mid \mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}), \theta] &= \frac{\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)}{\text{Prob}\,[\mathbf{S}(\mathbf{V}) = \mathbf{s}(\mathbf{v}) \mid \theta]} \\
&= \frac{\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)}{\sum_{\mathbf{s}(\mathbf{v}^*)=\mathbf{s}(\mathbf{v})} \text{Prob}\,(\mathbf{V} = \mathbf{v}^* \mid \theta)} \\
&= \frac{g[\mathbf{s}(\mathbf{v}), \theta]h(\mathbf{v})}{\sum_{\mathbf{s}(\mathbf{v}^*)=\mathbf{s}(\mathbf{v})} g[\mathbf{s}(\mathbf{v}), \theta]h(\mathbf{v}^*)} \\
&= \frac{h(\mathbf{v})}{\sum_{\mathbf{s}(\mathbf{v}^*)=\mathbf{s}(\mathbf{v})} h(\mathbf{v}^*)},
\end{aligned}$$

which is independent of θ, and we see that $\mathbf{S}(\mathbf{V})$ is therefore sufficient. *Thus (18.2.1) is a convenient necessary and sufficient condition for the sufficiency of a statistic.*

The following simple lemmas will be of use.

Lemma 18.2.1. A given statistic is sufficient when the parameter is restricted to any range consisting of an *arbitrary* pair of points if and only if it is sufficient when the parameter is unrestricted.

Proof. We note that the condition "no variation of a conditional probability as θ varies unrestrictedly" is equivalent to the condition "no variation of a conditional probability as θ varies over the values θ_1, θ_2" considered respectively for all pairs θ_1, θ_2 in the range of θ. □

Lemma 18.2.2. If a given statistic **S** is sufficient, and is minimal sufficient when the range of θ is restricted to two specified points θ_1, θ_2, then it is minimal sufficient.

Proof. Note that if **s** were not minimal sufficient, it would not be possible to write it as a function of some statistic **z**, which is sufficient for θ unrestricted. However, by Lemma 18.2.1, **z** is also sufficient for $\theta = \theta_1, \theta_2$, contradicting the minimal sufficiency of **s** under this restriction. □

18.3 Test Scores and Sufficient Statistics

We shall give examples of sufficient statistics based on response patterns for the following test models:

1. *Tests with equivalent items.* The items of such tests have a common item characteristic curve $P_1(\theta)$ of any form

$$\begin{aligned}\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) &= \prod_{g=1}^{n} P_1(\theta)^{u_g} Q_1(\theta)^{1-u_g} \\ &= P_1(\theta)^{\sum_{g=1}^{n} u_g}\, Q_1(\theta)^{n-\sum_{g=1}^{n} u_g} = P_1(\theta)^x Q_1(\theta)^{n-x} \\ &= [P_1(\theta)/Q_1(\theta)]^x Q_1(\theta)^n, \end{aligned} \tag{18.3.1}$$

where $x = \sum_{g=1}^{n} u_g$, denoting the number correct, is the sufficient statistic. This is an example of the general factored form (18.2.1), with the factor $h(\mathbf{v})$ taking the trivial but admissible form $h(\mathbf{v}) \equiv 1$.

2. *Logistic test model.* From (17.2.7), we have

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \left[\prod_{g=1}^{n} Q_g(\theta)\right] \exp\left(\theta D \sum_{g=1}^{n} a_g u_g\right) \exp\left(-D \sum_{g=1}^{n} a_g b_g u_g\right). \tag{18.3.2}$$

In this factored form we can recognize the weighted-sum score formula

$$x = \sum_{g=1}^{n} a_g u_g \tag{18.3.3}$$

as a sufficient statistic.

3. *Two-point discrimination problems with any test model.* It is interesting and useful to discuss an oversimplified version of the problem of classification into two ability levels in terms of a test model that may have the general form

$$\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}, \tag{18.3.4}$$

but whose parameter space is (unrealistically) restricted to consist of just two points, $\theta = \theta_1$ or θ_2 $(\theta_2 > \theta_1)$. We can then denote the model by

Prob $(\mathbf{V} = \mathbf{v} \mid \theta_\alpha)$, with $\alpha = 1$ or 2 now playing the role of the parameter. Thus

$$\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_\alpha) = \begin{cases} \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1) & \text{if } \alpha = 1, \\ \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_2) & \text{if } \alpha = 2, \end{cases}$$

or

$$\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_\alpha) = \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1) \left[\frac{\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_2)}{\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1)}\right]^{\alpha-1}, \qquad \alpha = 1 \text{ or } 2. \tag{18.3.5}$$

Letting $L(\mathbf{v}) = \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_2)/\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1)$, we have

$$\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_\alpha) = \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1) L(\mathbf{v})^{\alpha-1}, \qquad \alpha = 1 \text{ or } 2. \tag{18.3.6}$$

Here the factor depending on the parameter α is independent of $\mathbf{v}$ except through $L(\mathbf{v})$, the *likelihood ratio statistic*, which is therefore a sufficient statistic. We have

$$\begin{aligned} \log L(\mathbf{v}) &= \log \left[\frac{\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_2)}{\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1)}\right] \\ &= \log \prod_{g=1}^{n} \left[\frac{Q_g(\theta_2)}{Q_g(\theta_1)}\right] + \log \prod_{g=1}^{n} \left[\frac{P_g(\theta_2) Q_g(\theta_1)}{P_g(\theta_1) Q_g(\theta_2)}\right]^{u_g} \\ &= K + \sum_{g=1}^{n} w_g u_g, \end{aligned} \tag{18.3.7}$$

where K is independent of $\mathbf{v}$ and

$$w_g = w_g(\theta_1, \theta_2) = \log \left[\frac{P_g(\theta_2) Q_g(\theta_1)}{P_g(\theta_1) Q_g(\theta_2)}\right], \qquad g = 1, \ldots, k. \tag{18.3.8}$$

Thus the weighted-sum test score formula

$$x(\mathbf{v}) = \sum_{g=1}^{n} w_g u_g,$$

with *best weights* w_g defined by (18.3.8), is also a sufficient statistic.

We can readily verify that $L(\mathbf{v})$ is in fact the minimal sufficient statistic as defined at the end of Section 18.1. It suffices to use (18.3.6) to write the usual formula for the probability of one value of $L(\mathbf{v})$, conditionally on the occurrence of either this or one other value; and to observe that this conditional probability depends on that parameter. It follows that no statistic taking the same value over the two sets of points $\mathbf{v}$ on which $L(\mathbf{v})$ takes distinct values can be sufficient.

As we noted at the end of Section 18.1, if $L(\mathbf{v})$ is minimal sufficient, then so is any one-to-one transformation of it, such as $\log L(\mathbf{v}) - K = \sum_{g=1}^{n} w_g u_g$.

In the logistic case, we find that $w_g(\theta_1, \theta_2) = Da_g(\theta_2 - \theta_1)$, or

$$\sum_{g=1}^{n} w_g u_g = D(\theta_2 - \theta_1) \sum_{g=1}^{n} a_g u_g.$$

When D, θ_1, and θ_2 are fixed, this is a one-to-one transformation of $x = \sum_{g=1}^{n} a_g u_g$, the sufficient statistic considered in (2) above. This means that x is a minimal-sufficient statistic when θ is restricted to the values θ_1 and θ_2, and is sufficient for unrestricted θ. Since these are the conditions for Lemma 18.2.2, we may conclude that

$$x = \sum_{g=1}^{n} a_g u_g$$

is *minimal sufficient for the logistic model* when the range of θ is unrestricted.

18.4 Sufficiency and the Logistic Test Model

In this section, we shall demonstrate a kind of converse to the result just shown: We shall show that *only* the class of logistic test models admits weighted-sum statistics that are minimal sufficient whether the range of θ is restricted or unrestricted.

Finding a weighted-sum statistic that is merely sufficient is not significant. To illustrate this, consider, for example, the rather odd but possible weights

$$w_1 = 0.1, \quad w_2 = 0.01, \ldots, w_g = 10^{-g}, \ldots .$$

These determine a single weighted-sum scoring formula that is sufficient for *every* test model! To verify this, we need only observe that we can uniquely determine the response pattern $\mathbf{v}$ from any given value of the statistic

$$x(\mathbf{v}) \equiv \sum_{g=1}^{n} w_g u_g = \sum_{g=1}^{n} 10^{-g} u_g.$$

If, for example, $x(u_1, \ldots, u_n) = 0.101$, we can *deduce* that $u_1 = 1$, $u_2 = 0$, and $u_3 = 1$; and if n exceeds 3, then $u_g = 0$ for all $g > 3$. In fact these weights are devised so that x is just the decimal fraction with respective digits $u_1, u_2, \ldots$.

The characterization of the general logistic model given by the theorem and corollaries below is related to conditions, on more general classes of models, for existence of sufficient statistics having certain simple properties (see the discussions of Lindgren, 1962, p. 201, and the general exponential class in Lehmann, 1959, p. 51).

We must appreciate the role that the possible transformations of the ability scale discussed in Section 17.6 may play in any interpretation of this characterization of the logistic model. For example, each test with equivalent items can be given the logistic form by a rescaling of ability: In the example of $P_1(\theta)$

having the three-parameter logistic form (which is *not* a logistic form), we have

$$P_1(-\infty) = C_1 > 0,$$

and

$$P_1(\theta) > C_1 \quad \text{for all} \quad \theta > -\infty.$$

Solving the relation $\Psi(\theta^*) = P_1(\theta)$ for θ^*, we obtain

$$\theta^* = \theta^*(\theta) = \log [P_1(\theta)/Q_1(\theta)]$$

as a rescaling transformation that gives $P_1^*(\theta^*) = \Psi(\theta^*)$. The new θ^* ability scale necessarily has a lower bound θ^*, determined by $P_1^*(\theta^*) = \Psi(\theta^*) = C_1$; this is *entailed* by the original assumption of the ICC form, in which $P_1(\theta) > C_1$ for all $\theta > -\infty$.

The condition that a model has a logistic form, allowing for possible rescaling of ability, may be expressed thus:

There exists an ability scale $\theta^* = \theta^*(\theta)$ on which the ICCs $P_g(\theta)$ of items assume the logistic form

$$P_g^*(\theta^*) \equiv \Psi(A_g\theta^* + B_g) = P_g(\theta), \qquad g = 1, \ldots, k,$$

where A_g and B_g are constants.

The scale transformation required may be written

$$\theta^*(\theta) = A \log \frac{P_h(\theta)}{Q_h(\theta)} + B,$$

where $P_h(\theta)$ is the ICC of any selected item, and $A = 1/A_h$ and $B = -B_h/A_h$ are any constants, $A > 0$. We conclude that a model is logistic, up to possible rescaling of θ, if and only if there exists a transformation $\theta^*(\theta)$ such that for each item g, $\log [P_g(\theta)/Q_g(\theta)]$ is a linear function of θ^*.

Theorem 18.4.1. For $g = 1, 2$, let the $P_g(\theta)$ be any continuous strictly increasing functions defined for $\theta' < \theta < \theta''$, with $0 < P_g(\theta) < 1$. Let θ_1 be any fixed value of θ in the interval, and let

$$R(\theta) = w_2(\theta_1, \theta)/w_1(\theta_1, \theta) \qquad \text{for} \quad \theta \neq \theta_1,$$

where

$$w_g(\theta_1, \theta) = v_g(\theta) - v_g(\theta_1) \qquad \text{and} \qquad v_g(\theta) = \log [P_g(\theta)/Q_g(\theta)], \qquad g = 1, 2.$$

Then, if $R(\theta)$ is independent of θ, there exists a strictly increasing continuous function $\theta^* = \theta^*(\theta)$ such that $P_g^*(\theta^*) \equiv P_g(\theta)$ has the logistic form

$$P_g^*(\theta^*) = \Psi(A_g\theta^* + B_g), \qquad g = 1, 2,$$

where A_g and B_g are constants.

Proof. Let

$$\theta^*(\theta) = w_1(\theta_1, \theta).$$

This is a continuous, strictly increasing function of θ. Then we have the required form

$$P_1^*(\theta^*) = \Psi(A_1\theta^* + B_1),$$

with

$$A_1 = 1, \qquad B_1 = v_1(\theta_1) \equiv \log [P_1(\theta_1)/Q_1(\theta_1)].$$

Let $A_2 = R(\theta)$, by assumption a constant for $\theta \neq \theta_1$, and let $B_2 = v_2(\theta_1)$. Then

$$R(\theta) = \frac{v_2(\theta) - v_2(\theta_1)}{w_1(\theta_1, \theta)} = \frac{v_2(\theta) - B_2}{\theta^*} \equiv A_2,$$

or

$$A_2\theta^* + B_2 = v_2(\theta) = \log [P_2(\theta)/Q_2(\theta)]$$

for all values of θ. Thus $P_2^*(\theta^*)$ also has the required form

$$P_2^*(\theta^*) = \Psi(A_2\theta^* + B_2).$$

For $\theta = \theta_1$, we may write $A_2 w_1(\theta_1, \theta) = v_2(\theta) - v_2(\theta_1)$ and the result still holds. □

Corollary 18.4.2. Any test model (or any set of items) whose ICCs determine respective weights $w_g(\theta_1, \theta_2)$ for two-point problems, the ratios of which are constant as θ_1 and θ_2 vary ($\theta_1 \neq \theta_2$), is equivalent to a logistic test model (or a set of items with logistic ICCs).

Our final and sharpest result linking the logistic form with sufficiency is given as

Theorem 18.4.3. Suppose that two continuous ICCs $P_1(\theta)$, $P_2(\theta)$, give best weights for respective two-point problems, such that

$$R(\theta_1) = \frac{w_2(\theta_2, \theta_1)}{w_1(\theta_2, \theta_1)} < \frac{w_2(\theta_2, \theta_3)}{w_1(\theta_2, \theta_3)} = R(\theta_3),$$

where $\theta_1 < \theta_2 < \theta_3$. Then there is a test model whose items have ICCs only of these forms, in which

1) the minimal sufficient statistic for one two-point problem is not sufficient for a second two-point problem, and
2) the minimal sufficient statistic for the second problem is not minimal for the first.

Proof. We shall illustrate the method of proof by discussing a simple example of the hypothesis of the theorem. Suppose that

$$w_1(\theta_1, \theta_2) = w_2(\theta_1, \theta_2) = w_1(\theta_2, \theta_3) = 1, \qquad w_2(\theta_2, \theta_3) = 2.$$

Consider the test model with just two items, with ICCs $P_1(\theta)$ and $P_2(\theta)$, respectively. Then, first, the minimal sufficient statistic for the two-point problem concerning θ_1 and θ_2 is $s = s(u_1, u_2) = u_1 + u_2$, for which $s(1, 0) = s(0, 1) = 1$. Second, for the problem concerning θ_2 and θ_3, the minimal sufficient statistic is $t = t(u_1, u_2) = u_1 + 2u_2$, for which $t(1, 0) = 1 < t(0, 2) = 2$. Thus s is not sufficient for the second problem, and t is not minimal for the first. □

The property of minimal sufficiency is preserved under any one-to-one transformation of a statistic. Thus, if a weighted-sum scoring formula is transformed by a nonlinear one-to-one function, it will lose its weighted-sum form but retain any sufficiency or minimal sufficiency properties it may have had. If the transformation is nonmonotone, however, the particular simple ordering of response patterns $(u_1, \ldots, u_n)$ will usually also be altered. Furthermore, if the transformation has a multidimensional range, then no new (mathematically natural) simple ordering will replace the original one. Thus we see that in relation to sufficiency properties as such, considerations of specific forms of statistics, such as weighted-sums, have no essential substance. In our discussion above, it has therefore been considerations merely of simplicity and convenience that have led us to *choose* the particular form

$$\log L(\mathbf{v}) - K,$$

which is linear in the u_g, as a *representation* of the minimal sufficient statistic for any two-point problem. Thus sufficiency concepts alone do not include concepts of information about abilities which may be related to an ordering of response patterns $\mathbf{v}$, such as may be determined by any real-valued scoring formula $x(\mathbf{v})$. Such concepts are discussed, beginning with the treatment of classification rules in Chapter 19 below.

18.5 Sufficiency and the Information Structures of Tests

In Section 17.7, we interpreted the distributions $F(x \mid \theta)$ of test scores as representing the information structure of a test. We can now give additional theoretical and practical substance to this interpretation. If the test in question admits a real-valued sufficient statistic x, then the distributions $F(x \mid \theta)$ represent fully the information structure of a test in precisely the sense of the concept of sufficient statistics considered in the preceding sections.

In any case, the family of distributions $F(x \mid \theta)$ represents the information structure of the test in this practical sense: It characterizes the forms and error-probability properties of estimation and classification methods that can be based on the scoring formula x in the ways illustrated in Section 17.4. In our further development of estimation and classification methods, however, we must pay particular attention to cases in which there exists no real-valued sufficient statistic that has the relatively simple and convenient properties considered above.

References and Selected Readings

BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16.* Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January, 1957.

BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics,* 1958, **29,** 1284 (abstract). (c)

BIRNBAUM, A., Statistical theory of tests of a mental ability. *Annals of Mathematical Statistics,* 1958, **29,** 1285 (abstract). (d)

LEHMANN, E., *Testing statistical hypotheses.* New York: Wiley, 1959.

LINDGREN, B. W., *Statistical theory.* New York: Macmillan, 1960, 1962.

CHAPTER 19

CLASSIFICATION BY ABILITY LEVELS

19.1 Classification Rules for Distinguishing Two Levels of Ability

In this chapter, we shall consider prototype problems in which the information about ability levels in response patterns is used to order or classify examinees according to higher or lower ability. We shall develop some concepts and techniques which represent optimal solutions of such problems. These solutions in turn will provide information about the appropriate choice of score formulas for various applications. These solutions will also provide information on the dependence of these optimal solutions and score formulas on details of assumed underlying test models and details of specified purposes of classification of examinees by ability levels.

We shall consider the *efficiency* of classification rules of the form illustrated in Section 17.4, according to which an examinee's test score is used to determine his classification as *high ability* ($x > x_0$) or *low ability* ($x \leq x_0$). We shall also consider rules of more general form for such applications. "Efficiency" here refers to the error probability functions of such rules, which are illustrated in Fig. 17.4.3.

The natural basic meaning of "efficient" in such a context is "having adequately small probabilities for all relevant kinds of errors". In any situation where adequately small error probabilities of all relevant kinds are provided by an available test and classification rule, further considerations of efficiency may have little value. But in other situations there may be interest in reducing the error probabilities for given tests and in appraising, comparing, selecting, designing, or constructing new tests and classification rules. The more refined considerations of efficiency relevant here may be termed *statistical efficiency*. They arise when we attempt to use a test model to determine classification rules that are optimal in the sense that they minimize relevant error probabilities. They also arise in broader related contexts, including the design of test models for classification purposes.

The properties of statistical efficiency of classification rules and of estimators, and also the sufficiency property considered in the previous chapter, depend rather sensitively on the detailed form assumed for each item characteristic curve and on the product form by which the local independence assumption

is represented. Since in practice the forms of item characteristic curves are imperfectly known, it is of much practical and theoretical interest to determine to what extent a given inference method, derived under a given assumed form of test model, retains its relevant properties when somewhat different forms hold. It is also of interest to develop methods that are robust for relevant properties, that is, methods that are relatively safe in the sense that they are insensitive to variation in forms of models (within indicated limits). The development of *efficiency-robust* statistical inference methods has been undertaken systematically only in recent years, even for the most standard statistical problems, such as point and interval estimation of means. It has not yet been taken up systematically for models of tests such as those considered here; however, we shall discuss several points relevant to robustness of test models in Sections 19.3 and 20.7.

For any given test model that may be represented by a given probability function Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$, we can conveniently represent any definite rule for classifying examinees as high or low on the basis of their response patterns $\mathbf{v}$ as a function $d = d(\mathbf{v})$, taking the values one (high) and zero (low). Then, for examinees of any given ability level θ, we have

$$\begin{aligned} \text{Prob (high with rule } d \mid \theta) &\equiv \text{Prob } [d(\mathbf{V}) = 1 \mid \theta] \\ &= \sum_{d(\mathbf{v})=1} \text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta). \end{aligned}$$

The last expression can also be written and interpreted as $\mathscr{E}[d(\mathbf{V}) \mid \theta]$, the expected value of $d(\mathbf{V})$ at level θ. For any rules in which, for some score formula $x(\mathbf{v})$, $d(\mathbf{v}) = 1$ only when $x(\mathbf{v}) > x_0$, we have

$$\mathscr{E}[d(\mathbf{V}) \mid \theta] = 1 - F(x_0 \mid \theta),$$

where $F(x_0 \mid \theta) = \text{Prob } \{x(\mathbf{V}) \leqq x_0\}$.

19.2 Two-Point Classification Problems

Consider any test model and any classification rule $d(\mathbf{v})$, and let θ_1 be any specified ability level definitely considered low in a given context of application (for example, the largest such level). Then

$$\mathscr{E}[d(\mathbf{V}) \mid \theta_1] = \text{Prob } [d(\mathbf{V}) = 1 \mid \theta_1]$$

is one of the relevant error probabilities that should have a suitably small value, since it is the probability of a person of a certain low ability θ_1 being classified as high. Similarly, if θ_2 is any level considered definitely high (for example, the smallest such level), then

$$1 - \mathscr{E}[d(\mathbf{V}) \mid \theta_2] = \text{Prob } [d(\mathbf{V}) = 0 \mid \theta_2]$$

is another such relevant error probability. We can denote these, respectively, by

$$\alpha \equiv \alpha(d \mid \theta_1) \equiv \mathscr{E}[d(\mathbf{V}) \mid \theta_1] = \text{Prob}\,[d(\mathbf{V}) = 1 \mid \theta_1],$$
$$\beta \equiv \beta(d \mid \theta_2) \equiv 1 - \mathscr{E}[d(\mathbf{V}) \mid \theta_2] = \text{Prob}\,[d(\mathbf{V}) = 0 \mid \theta_2]. \qquad (19.2.1)$$

We may achieve a minimal $\alpha(d \mid \theta_1)$ for any test by adopting the rule that $d(\mathbf{V}) \equiv 0$. In fact this gives $\alpha = 0$, since this trivial rule never classifies anyone as high and therefore never misclassifies an examinee with low ability θ_1. But usually, of course, this ideal minimization of α is achieved only along with the unacceptable value $\beta = \beta(d \mid \theta_2) = 1$, since the rule described misclassifies each examinee of high level θ_2. Similarly $\beta = 0$ is attainable but usually only along with $\alpha = 1$. Thus the rules ordinarily of interest allow small positive probabilities of errors of each kind; among these, however, rules that minimize these probabilities jointly are to be preferred.

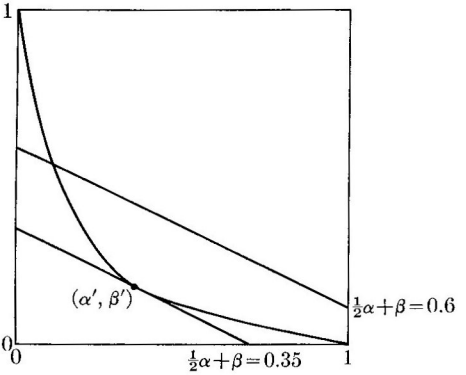

FIG. 19.2.1. $H(\alpha, \beta) = \frac{1}{2}\alpha + \beta$. The decision rule that yields values $\alpha = \alpha'$, $\beta = \beta'$, $\frac{1}{2}\alpha + \beta = 0.35$ is the rule that minimizes $H(\alpha, \beta)$. This is demonstrated by the fact that the $H(\alpha, \beta) = 0.35$ is tangent to the (α, β)-error curve.

We can usefully relate the general goal of determining $d(\mathbf{v})$ so as to jointly minimize $\alpha(d \mid \theta_1)$ and $\beta(d \mid \theta_2)$ to the mathematical problem represented in Fig. 19.2.1. Let A and B be any given positive numbers, and let

$$H = H(\alpha, \beta) = A\alpha + B\beta = A\alpha(d \mid \theta_1) + B\beta(d \mid \theta_2). \qquad (19.2.2)$$

This function is strictly increasing in each of the two error probabilities α and β of interest. Hence the problem of determining the form of $d(\mathbf{v})$ so as to minimize H seems at least qualitatively appropriate and relevant to our general goal, and we shall see that it leads to a useful mathematical formulation of our general problem and to a solution for it.* It is not difficult to show that $H(\alpha, \beta)$

* This development is well known in modern elementary mathematical statistics (see, for example, Birnbaum and Maxwell, 1960, pp. 157–159). It includes the Neyman-Pearson lemma (Lindgren, 1962, p. 238) and the derivation of the class of admissible tests between two simple hypotheses (Lindgren, 1962, pp. 162–163).

is minimized by $d(\mathbf{v})$ if and only if

$$d(\mathbf{v}) = \begin{cases} 1 & \text{if} \quad L(\mathbf{v}) \equiv \dfrac{\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta_2)}{\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta_1)} > \dfrac{A}{B}, \\ 0 & \text{if} \quad L(\mathbf{v}) < \dfrac{A}{B}. \end{cases} \tag{19.2.3}$$

(See, for example, Birnbaum and Maxwell, 1960, pp. 157–158.) The reader should note that if $L(\mathbf{v}) = A/B$ for some $\mathbf{v}$, then either value can be assigned to $d(\mathbf{v})$ without affecting the value of $H(\alpha, \beta)$. Such rules depend only on the statistic $L(\mathbf{v})$, the likelihood ratio statistic. In Section 18.3, we demonstrated that $\mathrm{L}(\mathbf{v})$ is a sufficient statistic only when the range of θ is restricted to the values θ_1, θ_2 only.

We might have specified the restriction $B \equiv 1 - A$, $0 < A < 1$, without restricting the range of A/B. Then H could be formally interpreted as a weighted average of the error probabilities α and β. In a Bayesian approach to inference, A and $1 - A$ could then be formally interpreted as respective prior probabilities of the abilities θ_1 and θ_2; and the minimization of H, as minimization of the total probability of error.

We may state a basic property of any rule $d(\mathbf{v})$ that minimizes such a function H, defined by any choice of A and B: No other rule $d^* = d^*(\mathbf{v})$ can have a smaller error probability of one kind unless it has a larger error probability of the other kind. This property of such rules is called *admissibility*. The proof of the stated property follows immediately: If H is minimized by $d(\mathbf{v})$, and $d^*(\mathbf{v})$ gives, say, $\alpha(d^* \mid \theta_1) < \alpha(d \mid \theta_1)$, then if $\beta(d^* \mid \theta_2) \leq \beta(d \mid \theta_2)$, we have

$$A\alpha(d^* \mid \theta_1) + B\beta(d^* \mid \theta_2) < A\alpha(d \mid \theta_1) + B\beta(d \mid \theta_2),$$

which, by assumption, is a minimum of H; but this is a contradiction. We may note that the term "admissible" is also applied to other statistical methods with respect to other efficiency criteria. In this broader context, an admissible method is one whose efficiency cannot be improved for one value of θ without some corresponding loss for another value of interest.

In the present context, any admissible rule d is also a *best* (or *most powerful*) *rule* of level $\alpha = \alpha(d \mid \theta_1)$, in that β is minimized for fixed α. Since the admissibility property is symmetric in θ_1 and θ_2, it is clear that an analogous term can be applied to d in relation to $\beta = \beta(d \mid \theta_2)$.

As A/B increases, fewer points satisfy $L(\mathbf{v}) \geq A/B$, and α decreases while β increases, according to the definition of $d(\mathbf{v})$ in (19.2.3). In many cases of interest, these successive increments are very small, affording a rather fine choice of α levels. It can be shown that each α can be realized exactly by an admissible rule if "randomized" rules are allowed (Lindgren, 1962, pp. 240–241).

Each rule $d(\mathbf{v})$ may be represented conveniently by a point

$$(\alpha, \beta) = [\alpha(d \mid \theta_1), \beta(d \mid \theta_2)]$$

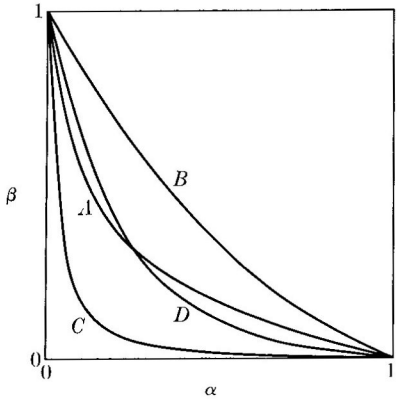

FIG. 19.2.2. Four hypothetical (α, β) curves.

in the unit square, as in the schematic Fig. 19.2.2. For any given test model, the admissible rules are thus represented by respective (α, β)-points, which, it can be shown, constitute a convex curve such as A, Fig. 19.2.2 (Lindgren, 1962, pp. 232ff). Rules essentially different from the admissible ones have error-probability points falling in the region strictly above curve A. For example, if the model is such that the best weights $w_g(\theta_1, \theta_2)$ for the items g are not all equal, then the classification rules based on the unweighted score $x(\mathbf{v}) = \sum_{g=1}^{n} u_g$, classifying high when its value exceeds a given constant x_0, would be inadmissible; the weights would be represented, respectively, by the (α, β)-points of a curve such as B. Rules represented by the (α, β)-points of C, and those points of D that fall below A, do not exist in connection with the test model for which curve A represents the admissible rules, because their existence would contradict the admissibility of those rules. Only by increasing the number of items in the given test model, or by adopting an otherwise different test model, can rules be attained that are superior to those of curve A.

Let

$$x = x(\mathbf{v}) = \log L(\mathbf{v}) - K = \sum_{g=1}^{n} w_g u_g,$$

where

$$w_g = v_g(\theta_2) - v_g(\theta_1) \qquad \text{and} \qquad v_g(\theta) = \log [P_g(\theta)/Q_g(\theta)],$$

as in Section 18.4. We note that x is an increasing function of $L(\mathbf{v})$, and from this we see that for two-point problems, with any test model, the admissible rules include all those that have the simple form in which high ability is indicated just by $x(\mathbf{v}) \geq x_0$ for some number x_0. The other admissible rules differ from this form only in the ways in which they are defined for the score value $x(\mathbf{v}) = x_0$. Thus a rule that indicates high ability just when $x(\mathbf{v}) > x_0$ is also admissible, and so is a rule that differs from this only by having a more

complicated form (possibly depending on $\mathbf{v}$ or on an auxiliary randomization variable) just when $x(\mathbf{v}) = x_0$. If each single value of $x(\mathbf{v})$, such as x_0, has a very small probability under each θ, as is true with many models, then such variations in the definition of $d(\mathbf{v})$ give alternative admissible rules which are nearly identical in their error-probability properties.

The two-point problem is of interest primarily as a prototype and as a technical step toward more realistic analyses in which all abilities definitely considered either high or low are taken into account. Figure 17.4.3 shows the error probabilities of a rule of the form illustrated above, when the full range of θ is considered. Rules having the admissibility property in two-point problems often can be proved admissible also where the range of θ is unrestricted. In this situation, an admissible rule is one such that no other rule can have a smaller error probability at one value of θ in the range $\theta \leq \theta_1$ or $\theta \geq \theta_2$ without having a larger error probability at some other value in the same range.

As an illustration, suppose that we have the problem of classification as high ability, $\theta \geq \theta_2$, or low ability, $\theta \leq \theta_1 (\theta_1 < \theta_2)$. Consider a rule $d(\mathbf{v})$ that is admissible for the two-point problem, $\theta = \theta_1$ or $\theta = \theta_2$, and that classifies as high just when $x(\mathbf{v}) \equiv \log L(\mathbf{v}) - K > x_0$, with positive error probabilities $\alpha = \alpha(d \mid \theta_1)$, $\beta = \beta(d \mid \theta_2)$, where $\alpha = 0.03$, $\beta = 0.08$, as in Section 17.4.3. Suppose further that x_0 is not among the attainable values of $x(\mathbf{v})$. Then ambiguities regarding classification when $x(\mathbf{v}) = x_0$ disappear, and we see that $d(\mathbf{v})$ is the *unique* rule that minimizes (19.2.2) for the A and B corresponding to x_0. This, in turn, implies that $d(\mathbf{v})$ is the only rule with error probabilities α and β at θ_1 and θ_2, respectively. Since $d(\mathbf{v})$ is admissible for this two-point problem, we may now conclude that it is admissible in the more general case: Any $d^*(\mathbf{v})$ with a smaller error probability than $d(\mathbf{v})$ at some $\theta' (\theta' \leq \theta_1$ or $\theta' \geq \theta_2)$ must have a larger error probability at either θ_1 or θ_2.

Two points, such as $[\theta_1, \alpha(d \mid \theta_1)]$, $[\theta_2, \beta(d \mid \theta_2)]$, on the error probability curve of any admissible rule provide a useful, simple partial description of that rule's properties. At the same time, two such points represent bounds on attainable error probabilities in the sense that these values α, β cannot be simultaneously improved on by any other rule based on the same mental test.

The scoring formula considered in the preceding comments was suggested by its admissibility for the two-point problem. Somewhat different choices of values θ_1, θ_2, chosen to mark a *highest definitely low ability* and a *lowest definitely high ability* schematically, would in general determine somewhat different weights w_θ. Of course the admissibility (or best) property and the results of the preceding discussion will hold again for each such rule.

We shall next consider the question: What is the most general test model for which the *same* weighted-sum scoring formula is obtained in *each* such determination of an admissible rule? Another form of this question is: As θ_1, θ_2 vary (subject to $\theta_1 < \theta_2$), what is the most general test model (if one exists) that admits a single classification rule that is best for all values $\theta_2 > \theta_1$, a so-called *uniformly best classification rule?* The answer is: The logistic test

model. The proof is based on considerations like those of Section 18.4, and on the observation that when θ_1 is fixed and θ_2 varies, if one scoring formula is to remain identically equal to

$$\log L(\mathbf{v}) \equiv \log [\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_2)/\text{Prob } (\mathbf{V} = \mathbf{v} \mid \theta_1)]$$

[apart from a factor $K(\theta_2)$, which may vary with θ_2 but is independent of $\mathbf{v}$], then the ratios of weights w_g/w_h in the scoring formula must be independent of variation in θ_2.

We have seen here that the two plausible concepts, sufficiency and admissibility, tend to agree with, and in a sense to complement, one another by supporting from somewhat different viewpoints the use of certain statistics (score formulas). Each concept seems to derive further plausibility from appearing compatible with, and complementary to, the other plausible concept. This in turn seems to confirm the plausibility and appropriateness of a general approach to statistical inference problems (including classification, estimation, and testing hypotheses) that incorporates these as basic concepts. Indeed the statistical methods thus obtained seem highly appropriate in some contexts, such as some situations requiring acceptance or rejection of applicants. But it would be a misleading oversimplification to suggest that we have adequately illustrated the significance of these concepts for statistical methodology in general. In fact other aspects of sufficiency and of error probabilities, particularly as concepts of statistical evidence in empirical research contexts, are among the basic concepts of statistical inference currently undergoing critical reappraisal; the reader might see, for example, Savage (1962), Birnbaum (1962, 1968a, 1968b), Novick and Hall (1965), Tukey (1962), Mosteller and Wallace (1964), and Hartigan (1965).

19.3 Locally Best Weights and Classification Rules

With many models, the detailed determination of best weights, and of related statistical theory and calculations, can be carried through most conveniently and completely in a special limiting case, namely, the case in which the difference $\theta_2 - \theta_1$ becomes arbitrarily small. Some detailed treatment of this case will provide convenient illustrations here. Even more importantly, it will be convenient to base the principal methods of estimation to be developed in the next chapter on the techniques developed here.

For θ_2 sufficiently near θ_1, the best weights $w_g = w_g(\theta_1, \theta_2)$ are effectively and conveniently approximated by the *locally best weights*, defined by

$$w_g \equiv w_g(\theta_1) = \lim_{\theta_2 \to \theta_1} [w_g(\theta_1, \theta_2)/(\theta_2 - \theta_1)]. \tag{19.3.1}$$

(We note that division of each weight $w_g(\theta_1, \theta_2)$ by the same positive constant $(\theta_2 - \theta_1)$ does not change the ratios between weights of respective items; just

these ratios characterize best weights.) Since

$$\frac{w_g(\theta_1, \theta_2)}{\theta_2 - \theta_1} = \frac{v_g(\theta_2) - v_g(\theta_1)}{\theta_2 - \theta_1}, \tag{19.3.2}$$

we have

$$w_g(\theta_1) = \lim_{\theta_2 \to \theta_1} \frac{v_g(\theta_2) - v_g(\theta_1)}{\theta_2 - \theta_1} = \frac{\partial}{\partial \theta} v_g(\theta)\Big|_{\theta=\theta_1}$$

$$= \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)}\Big|_{\theta=\theta_1} = \frac{P'_g(\theta_1)}{P_g(\theta_1)Q_g(\theta_1)}, \tag{19.3.3}$$

where $P'_g(\theta) = (\partial/\partial\theta)P_g(\theta)$, assuming that the derivative exists (as it does in most of our examples). To see that ratios between these weights for items g and h are close approximations to ratios between best weights $w_g(\theta_1, \theta_2)$ and $w_h(\theta_1, \theta_2)$ when θ_2 is sufficiently near θ_1, observe that when $\theta_2 - \theta_1$ is sufficiently small, we have (by the definition of a derivative)

$$w_g(\theta_1, \theta_2) \doteq \frac{\partial}{\partial \theta_2} w_g(\theta_1, \theta_2)\Big|_{\theta_2=\theta_1} (\theta_2 - \theta_1) \equiv w_g(\theta_1)(\theta_2 - \theta_1),$$

and hence

$$\frac{w_g(\theta_1, \theta_2)}{w_h(\theta_1, \theta_2)} \doteq \frac{w_g(\theta_1)}{w_h(\theta_1)}$$

if $w_h(\theta_1) \neq 0$ and $w_h(\theta_1, \theta_2) \neq 0$.

It can also be shown, as we might expect, that a classification rule based on a score t with locally best weights $w_g(\theta_1)$ has, of all rules with the same value of $\alpha = \alpha(d \mid \theta_1) = 1 - F(t_0 \mid \theta_1)$, the largest possible value of the derivative at θ_1 of the error-probability function

$$\frac{\partial}{\partial \theta}[1 - F(t_0 \mid \theta)]\Big|_{\theta=\theta_1} = -F'(t_0 \mid \theta_1),$$

provided the derivatives $P'_g(\theta)$ of the ICCs exist. Such a rule is called *locally best* (at θ_1); the reader might see, for example, Lehmann (1959, p. 364) and Rao (1965, pp. 382–383). The maximum condition means that among all rules with the given error probability α at θ_1, for each positive ϵ, no other rule has error probabilities β at least equally small for all θ_2 in the interval

$$\theta_1 < \theta_2 < \theta_1 + \epsilon.$$

We indicated near the end of Section 19.2 that if a test model has best weights $w_g(\theta_1, \theta_2)$, which are independent of θ_1 and θ_2 (at least as regards their mutual ratios), then the model is logistic. Now locally best weights $w_g(\theta_1)$ are essentially (that is, in their ratios) limits as $\theta_2 \to \theta_1$ of weights $w_g(\theta_1, \theta_2)$; and it can be shown that such limits are essentially independent of θ_1 only if the $w_g(\theta_1, \theta_2)$ are essentially independent of θ_1 and θ_2. Thus a test model has locally best

weights independent of θ_1 if and only if it is a logistic model. Examples of locally best weights are given below.

1. Logistic model

$$w_g(\theta) = Da_g \frac{\psi[DL_g(\theta)]}{\Psi[DL_g(\theta)]} \{1 - \Psi[DL_g(\theta)]\} = Da_g.$$

The final expression is easily obtained using (19.3.3). On canceling the inessential common factor D of such weights, we obtain the weights a_g, which have become familiar in other connections, in particular as best weights $w_g(\theta_1, \theta_2) \equiv a_g$, independent of θ_1 and θ_2.

2. Normal ogive model

$$w_g(\theta) = a_g \frac{\varphi[L_g(\theta)]}{\Phi[L_g(\theta)]\Phi[-L_g(\theta)]}.$$

The factor multiplying a_g here is the function $J(s) = \varphi(s)/\Phi(s)\Phi(-s)$, where $s = L_g(\theta) = a_g(\theta - b_g)$ depends on $\theta - b_g$ as well as on a_g. The behavior of $J(s)$ is indicated sufficiently for our purposes by a brief table:

s	0	± 1	± 2	± 3
$J(s)$	1.6	1.8	2.4	3.4

Thus, for efficient discrimination between ability levels in the neighborhood of a given level θ_1, the weighting w_g to be given to response u_g will vary by a factor possibly as large as 2, depending upon the item's difficulty b_g (through the difference $\theta_1 - b_g$).

One starting point for an investigation of robustness of validity and efficiency properties would be a study of error-probability functions of classification rules that are optimal under certain logistic test models when normal-ogive models with corresponding item-parameters are in fact valid; and conversely.

3. Three-parameter logistic model

$$w_g(\theta) = Da_g\Psi[DL_g(\theta) - \log c_g].$$

The moments and distributions of composite scores with such weights were discussed in Section 17.7; and applications of such composite scores are discussed throughout Chapters 17 through 21.

19.4 More General Classification Rules, Composite Scores, and Statistical Efficiency in General

The considerations of the preceding sections can be extended to problems of classification into three or more ranges of ability; *low*, *middle*, and *high*. Admissible rules for three-point problems, for example, are obtainable on the

basis of a direct generalization of the derivation of admissible two-point rules given above (see, for example, Birnbaum and Maxwell, 1961). No single best weighted score alone will determine all the admissible rules, except in logistic test models. In other models, at least a pair of such statistics jointly, or else a statistic of another form, must be used as the basis for admissible rules. For example, a score with weights $w_g(\theta_1, \theta_2)$, and another with weights $w_g(\theta_2, \theta_3)$, are jointly sufficient when θ is restricted to three values $\theta_1, \theta_2, \theta_3$ in any test model with $0 < P_g(\theta_i) < 1$ for all i, g.

It can be shown that the most general test model for which a single score formula suffices for such purposes is the logistic model. Rules having the simple form determined by two critical values x_1 and x_2, with scores $x(\mathbf{v}) \leq x_1$ classified as low, scores $x(\mathbf{v}) \leq x_2$ classified as high, and others classified as middle, are admissible in each case of the logistic model [when $x(\mathbf{v}) = \sum_{g=1}^{n} a_g u_g$]. But for essentially different models (with nonequivalent items), rules of this form, based on any score formula $x(\mathbf{v})$, are in general inadmissible. The method of proof is a direct extension of that for the two-point problem, which we indicated near the end of Section 19.2.

Another general mode of formulating inference or classification problems, partly related to formulations in terms of error probabilities, is that of statistical decision theory. After specifying the costs or disutilities of errors of each of the numerous kinds possible in a given problem, one can characterize admissible classification rules in such general problems. In problems involving a single real-valued parameter and ordered alternative decisions, it can be shown under certain broad conditions that only in the general logistic test model are all the admissible decision rules based on a single real-valued statistic whose increasing values indicate the respective ordered decisions (see Lindgren, 1962, pp. 205–209, and references therein).

A mathematical condition that characterizes and entails these simple, intuitively natural forms for admissible rules for a broad class of inference and decision problems is the *monotone likelihood ratio property*: For any fixed x_1, x_2 where $x_2 > x_1$,

$$\text{Prob}\,[x(\mathbf{V}) = x_2 \mid \theta]/\text{Prob}\,[x(\mathbf{V}) = x_1 \mid \theta]$$

is strictly increasing in θ. The logistic model and score formula meet this condition. Another closely related condition that holds with the logistic model is that Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ has the *exponential* (or Koopman-Darmois) form. The latter condition entails both the existence of a sufficient statistic of the weighted sum form and the monotone likelihood ratio property for the distributions of that statistic (see, for example, Lindgren, 1962, pp. 200–202, and references therein).

Rules of these intuitively natural forms are not in general admissible with models essentially different from the logistic. But most thinking about test scores, most practice with them, seems to incorporate, at least tacitly, the assumption that a single suitably defined composite score (usually the un-

weighted sum, occasionally a weighted sum) serves efficiently to indicate respective ordered decisions or inferences in a monotone way. The properties of sufficiency, monotonicity, and statistical efficiency discussed above may be regarded as supplying some *explications* of this tacit assumption. But on these terms, the assumption is given *justification* only under the very restrictive condition that a test model have the logistic form; with models of essentially different form, some loss of efficiency is entailed. We shall give detailed consideration to some quantitative aspects of such losses; for example, in Section 20.8 we shall appraise losses of precision in estimation based on nonoptimal score formulas.

The tacit assumption just mentioned is doubtless related to the plausible but overly simple expectation that efficient statistical classification rules should resemble efficient nonstatistical rules; in particular, that some estimate of an unknown value θ should visibly play the role in the classification procedure that would be played by θ if it were known. This point about test scores and their uses is in some respects akin to certain subtle questions of general methodology which have been of much interest in the development of mathematical statistics. The latter are basic questions in the logic of measurement in the presence of errors of measurement; they involve the intuitive and the operational aspects of statistical methods for a single unknown parameter, and the relations between efficient estimators, sufficient statistics, and such properties as monotone likelihood ratio and the Koopman-Darmois form.

19.5 Quantitative Appraisal and Efficient Design of Classification Rules

Let θ_1 and θ_2 be ability levels representing, respectively, specified low and high levels of ability as above. Consider any score formula of the weighted sum form $x(\mathbf{v}) = \sum_{g=1}^{n} w_g u_g$, and any classification rule of the form that classifies as high when $x > x_0$, and as low when $x \leq x_0$, where x_0 is a specified number (*critical score*). The error probability function of such a rule is

$$1 - F(x_0 \mid \theta) \quad \text{for} \quad \theta \leq \theta_1, \qquad \text{and} \qquad F(x_0 \mid \theta) \quad \text{for} \quad \theta \geq \theta_2,$$

as illustrated in Fig. 17.4.3. We shall focus attention particularly on the error probabilities

$$\alpha = \alpha(d \mid \theta_1) = 1 - F(x_0 \mid \theta_1) \quad \text{and} \quad \beta \equiv \beta(d \mid \theta_2) = F(x_0 \mid \theta_2), \tag{19.5.1}$$

where $d = d(\mathbf{v})$ refers to a particular rule of the form described. Here we shall ordinarily use the normal approximation presented in Section 17.7:

$$F(x_0 \mid \theta) \doteq \Phi\left\{\frac{x_0 - \mathscr{E}(x(\mathbf{V}) \mid \theta)}{\sigma[x(\mathbf{V}) \mid \theta]}\right\}. \tag{19.5.2}$$

For notational simplicity in this section, we shall henceforth replace the approximate equality symbol $\doteq$, where it would ordinarily occur with this normal

approximation, by the equality symbol $=$. The approximate equality symbol will be used for other relations where appropriate, unless the contrary is specifically stated.

On this basis it is convenient to deduce formally a number of results. These results will provide useful general guidance. Also, in specific problems of interest they will provide quantitative solutions which may be treated as first approximations. In applications these solutions can be considered in connection with appropriate bounds on errors of approximation that may be available, or with other checks.

On these terms, we have

$$\alpha = 1 - \Phi\left\{\frac{x_0 - \mathscr{E}[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_1]}\right\} \quad \text{and} \quad \beta = \Phi\left\{\frac{x_0 - \mathscr{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]}\right\}. \tag{19.5.3}$$

Now

$$\mathscr{E}[x(\mathbf{V}) \mid \theta] = \sum_{g=1}^{n} w_g P_g(\theta) \quad \text{and} \quad \operatorname{Var}[x(\mathbf{V}) \mid \theta] = \sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta). \tag{19.5.4}$$

When we substitute the formulas (19.5.4) in formulas (19.5.3), we obtain formulas (19.5.5) and (19.5.6), which give α and β explicitly in terms of the item parameters and which also give other specific details of the structure of our test model and classification rule:

$$\alpha = 1 - \Phi\left\{\frac{x_0 - \sum_{g=1}^{n} w_g P_g(\theta_1)}{\left[\sum_{g=1}^{n} w_g^2 P_g(\theta_1) Q_g(\theta_1)\right]^{1/2}}\right\}, \tag{19.5.5}$$

$$\beta = \Phi\left\{\frac{x_0 - \sum_{g=1}^{n} w_g P_g(\theta_2)}{\left[\sum_{g=1}^{n} w_g^2 P_g(\theta_2) Q_g(\theta_2)\right]^{1/2}}\right\}. \tag{19.5.6}$$

These relations are the basis for a number of deductions relevant to appraisal and design both of classification rules and of tests to be used with classification rules. At the same time, these considerations constitute a convenient technical step in developing estimation methods, and methods of appraisal of information structures of tests and design of tests, as we shall show in the next chapter.

Critical score. Suppose we are given any test model, represented by

$$\text{Prob}(\mathbf{V} = \mathbf{v} \mid \theta) = \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g}, \tag{19.5.7}$$

and any score formula $x = \sum_{g=1}^{n} w_g u_g$. Consider the problem of finding a

critical score x_0 that determines a classification rule $d = d(\mathbf{v})$ of the form in which $x(\mathbf{v}) > x_0$ classifies as high, and such that the error probability $\alpha(d \mid \theta_1)$ has a specified value α (for example, $\alpha = 0.06$) at a specified low ability θ_1. We require then that

$$\alpha = 1 - \Phi\left\{\frac{x_0 - \mathscr{E}[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_1]}\right\}. \tag{19.5.8}$$

This determines the required value

$$\begin{aligned} x_0 &= \mathscr{E}[x(\mathbf{V}) \mid \theta_1] + \Phi^{-1}(1 - \alpha)\sigma[x(\mathbf{V}) \mid \theta_1] \\ &= \sum_g^n w_g P_g(\theta_1) + \Phi^{-1}(1 - \alpha)\left[\sum_{g=1}^n w_g^2 P_g(\theta_1)Q_g(\theta_1)\right]^{1/2}. \end{aligned} \tag{19.5.9}$$

To determine the error probability of this rule at any specified high ability level θ_2, we substitute (19.5.9) in (19.5.6) to obtain

$$\begin{aligned} \beta &= \beta(d \mid \theta_2) \\ &= \Phi\left\{\frac{\mathscr{E}[x(\mathbf{V}) \mid \theta_1] + \Phi^{-1}(1 - \alpha)\sigma[x(\mathbf{V}) \mid \theta_1] - \mathscr{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]}\right\} \\ &= \Phi\left\{\Phi^{-1}(1 - \alpha)\frac{\sigma[x(\mathbf{V}) \mid \theta_1]}{\sigma[x(\mathbf{V}) \mid \theta_2]} + \frac{\mathscr{E}[x(\mathbf{V}) \mid \theta_1] - \mathscr{E}[x(\mathbf{V}) \mid \theta_2]}{\sigma[x(\mathbf{V}) \mid \theta_2]}\right\}, \end{aligned} \tag{19.5.10}$$

into which we can further substitute the summation formulas (19.5.4) for the moments of $x(\mathbf{V})$, when required.

Required number of items. In Sections 18.2 through 18.4, we have seen that for any test with equivalent items, the admissible rules are based on equal weighted composite scores $x(\mathbf{v}) = \sum_{g=1}^n u_g$. We have

$$\mathscr{E}[x(\mathbf{V}) \mid \theta] = nP(\theta), \qquad \sigma^2[x(\mathbf{V}) \mid \theta] = nP(\theta)Q(\theta), \tag{19.5.11}$$

and

$$\beta = \Phi\left\{\Phi^{-1}(1 - \alpha)\left[\frac{P(\theta_1)Q(\theta_1)}{P(\theta_2)Q(\theta_2)}\right]^{1/2} + n^{1/2}\frac{P(\theta_1) - P(\theta_2)}{[P(\theta_2)Q(\theta_2)]^{1/2}}\right\}. \tag{19.5.12}$$

If $P(\theta_1) - P(\theta_2) < 0$ when $\theta_2 > \theta_1$, as we have usually assumed (and have illustrated by most of the models of Chapter 17), we see that β decreases to zero as the number n of items increases, regardless of any more specific features of the common item characteristic curve $P(\theta)$.

A problem of test design that can be solved explicitly in this case is that of determining the minimal number n of items required to meet two given bounds α^* and β^* on error probabilities $\alpha(d \mid \theta_1) \leq \alpha^*$ and $\beta(d \mid \theta_2) \leq \beta^*$. Let us temporarily ignore the fact that n takes only integral values. We then see that

the last equation gives

$$n = \left\{\frac{\Phi^{-1}(1-\alpha^*)[P(\theta_1)Q(\theta_1)]^{1/2} - \Phi^{-1}(\beta^*)[P(\theta_2)Q(\theta_2)]}{[P(\theta_2) - P(\theta_1)]^2}\right\}^2. \quad (19.5.13)$$

The required number of items is the smallest integer greater than or equal to the preceding quantity.

When $\beta^* = \alpha^*$ has been specified, $\Phi^{-1}(\beta^*) = -\Phi^{-1}(1-\alpha^*)$. Then, on rearranging (19.5.13), we have

$$\frac{\Phi^{-1}(1-\alpha^*)}{n^{1/2}} = \frac{P(\theta_2) - P(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2} + [P(\theta_2)Q(\theta_2)]^{1/2}}. \quad (19.5.14)$$

The right-hand member of (19.5.14) depends just on the form of the common ICC and *constitutes a natural measure of the amount of information per item provided by items of the given form,* a measure of information specifically related to the contributions of items toward minimizing common error probabilities in discriminating between levels θ_1 and θ_2.

Local case: a measure of information. Of course, the right-hand member of (19.5.14) depends on θ_1 and θ_2. If we divide the right-hand member by $(\theta_2 - \theta_1)$, it becomes a kind of information measure expressed on the scale "per unit separation between ability levels". We may now obtain a particularly convenient approximation formula by proceeding to the limit $(\theta_2 \to \theta_1)$, provided that it exists, as indeed it does in most of the examples of ICCs in Chapter 17. This limit is

$$\frac{1}{2}\frac{P'(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2}}, \qquad \text{where} \quad P'(\theta) = \frac{\partial}{\partial\theta}P(\theta). \quad (19.5.15)$$

On squaring this expression and deleting the constant factor, we obtain the formulation

$$I(\theta_1, u_g) = \frac{P'(\theta_1)^2}{P(\theta_1)Q(\theta_1)} \quad (19.5.16)$$

as a measure of *information* per item having ICC of the form $P(\theta)$, that can be used to discriminate abilities in a neighborhood of θ_1. Using the limit approximation in (19.5.14), we have

$$\frac{\Phi^{-1}(1-\alpha^*)}{n^{1/2}} = \frac{P(\theta_2) - P(\theta_1)}{[P(\theta_1)Q(\theta_1)]^{1/2} + [P(\theta_2)Q(\theta_2)]^{1/2}}$$
$$\doteq 2\frac{P'(\theta_1)}{[2P(\theta_1)Q(\theta_1)]^{1/2}}(\theta_2 - \theta_1),$$

or

$$nI(\theta_1, u_g) \doteq 4\left[\frac{\Phi^{-1}(1-\alpha^*)}{(\theta_2 - \theta_1)}\right]^2. \quad (19.5.17)$$

This formula makes explicit the role of the item information function in characterizing the contribution, per item having given ICC, toward reduction of probabilities of errors concerning close alternative values θ_1 and θ_2. Thus α^* is lowered the same amount by doubling $I(\theta_1, u_g)$ as by doubling the number of items.

In the following chapter, we shall see that the *item information function* $I(\theta, u_g)$, represented by (19.5.16) taken as a function of θ, is related to the information functions $I(\theta, x)$ defined in Section 17.7; and that each of these plays a basic role in the theory and techniques of estimation.

Best difficulty levels. Let us consider an example of the uses of the general formulas (19.5.10) and (19.5.12) for β in problems of item selection and the design of tests and classification rules. Suppose that

1) θ_1, $\alpha = \alpha(d \mid \theta_1)$, n, and θ_2 ($\theta_2 > \theta_1$) are given, and the scale for θ is fixed;

2) we may choose any set of n items with *logistic* ICCs.

Our problem is to minimize $\beta(d \mid \theta_2)$ by appropriate choice of a logistic test design (model). In other words, we must choose a set of the item parameters $(a_1, b_1, a_2, b_2, \ldots, a_n, b_n)$, and a classification rule, that is, a critical score x_0, so as to minimize β. We recall (see Section 17.2) that a single logistic item with a large enough value of a_g would suffice to give arbitrarily small values to α and β, provided that b_g lies between θ_1 and θ_2. Experience with test items shows that there is an effective upper bound on the discriminating power of items ordinarily available. Hence we may add an upper bound on the a_g, say a', to our problem, and further simplify and schematize it for illustrative purposes by requiring that the items be equivalent. (It is plausible, and can be shown formally, that a best choice would in fact be one with each a_g equal to the given upper bound a' and with some common value b chosen for the b_g.) Thus reduced, our problem is to choose b and x_0 so as to minimixe β as given in (19.5.12).

Approximate determination of best difficulty level for a special case. If $\alpha = \alpha(\theta_1) = \frac{1}{2}$, then $\Phi^{-1}(1 - \alpha) = 0$, and we see from the form of (19.5.12) that β is minimized when

$$[P(\theta_2) - P(\theta_1)]/[P(\theta_2)Q(\theta_2)]^{1/2} \tag{19.5.18}$$

is maximized, or, equivalently, when

$$\frac{P(\theta_2) - P(\theta_1)}{\theta_2 - \theta_1} \frac{1}{[P(\theta_2)Q(\theta_2)]^{1/2}} \tag{19.5.19}$$

is maximized. As we saw above, the limit of this expression, as $\theta_2 \to \theta_1$, is

$$\tfrac{1}{2}\sqrt{I(\theta_1, u_1)}. \tag{19.5.20}$$

If θ_2 is sufficiently near θ_1, then the value b that maximizes $I(\theta_1, u_1)$ is an approximate solution of our problem. [Clearly this value of b also maximizes the slope of $F(x_0 \mid \theta)$ at $\theta = \theta_1$.]

In Section 20.4, we shall examine and illustrate the forms of item information curves for specific models. We shall show there that for both the logistic and normal ogive models, $I(\theta, u_g)$ is symmetric and unimodal in $\theta - b$, and hence $I(\theta_1, u_g)$ is maximized when $b = \theta_1$.

Exact determinations of best difficulty levels. It might be of interest to have explicit formulas or numerical tables for the value of b that minimizes the expression (19.5.12) for β, for given values of θ_1, θ_2, α, and n, and given forms of $P(\theta_1)$ and $P(\theta_2)$ in which b is the sole variable parameter. It might be convenient and of interest to combine this derivation with derivations for the similar problem in which θ_1, θ_2, α, β, and the forms of $P(\theta_1)$ and $P(\theta_2)$ are given, and it is required to determine (1) a value of b that meets (or exceeds) these requirements with the smallest possible number n of items and (2) to determine that minimum n. Such formulas are not now available, but the problems stated are readily solved in specific cases by successive numerical trials.

Exercises

19.1. Let $P(\theta) = \Psi[(1.7)(0.5)(\theta - b)]$, $\theta_1 = 1$, $\theta_2 = 2$, $\alpha = 0.01$, and $n = 49$. Compute β from (19.5.12) for the cases $b = 0, 1, 2, 3, 0.5, 1.5, 2.5$, and, if desired, for other successive trial values as well, to approximate a value of b that minimizes β.

19.2. Using the same specifications, but omitting the given value of n and adding the requirement that $\beta = 0.10$, determine approximately, by use of trial values of b, a value of b that minimizes the required value of n.

19.3. Give numerical examples relevant to the questions of robustness indicated in connection with the examples in Section 19.3.

References and Selected Readings

BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.

BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics,* 1958, **29,** 1284 (abstract). (c)

BIRNBAUM, A., Statistical theory of tests of a mental ability. *Ibid,* 1285 (abstract). (d)

BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics,* 1961, **32,** 112–137. (a)

BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)

BIRNBAUM, A., On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 1962, **57**, 269–326.

BIRNBAUM, A., Likelihood. *International Encyclopedia of the Social Sciences*, 1968. (a)

BIRNBAUM, A., Concepts of statistical evidence. In S. Morgenbesser, P. Suppes, and M. White (Eds.), *Essays in Honor of Ernest Nagel*. New York: St. Martin's Press, 1968. (b)

BIRNBAUM, A., and A. E. MAXWELL, Classification procedures based on Bayes' formula. *Applied Statistics*, 1960, **9**, 152–169. Reprinted in L. J. Cronbach and Goldine Gleser, *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press, 1965.

HARTIGAN, J. A., The asymptotically unbiased prior distribution. *Annals of Mathematical Statistics*, 1965, **36**, 1137–1152.

LEHMANN, E., *Testing statistical hypotheses*. New York: Wiley, 1959.

LINDGREN, B. W., *Statistical theory*. New York: Macmillan, 1960–1962.

LORD, F. M., A theory of test scores. *Psychometric Monograph, No. 7*. Chicago: University of Chicago Press, 1952. (a)

LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181–194. (b)

LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57–76.

MOSTELLER, F., and D. WALLACE, *Inference and disputed authorship: The Federalist*. Reading, Mass.: Addison-Wesley, 1964.

NOVICK, M. R., and W. J. HALL, A Bayesian indifference procedure. *Journal of the American Statistical Association*, 1965, **60**, 1104–1117.

RAO, C. R., *Linear statistical inference and its applications*. New York: Wiley, 1965.

SAVAGE, L. J., *The foundations of statistical inference*. New York: Wiley, 1962.

TUKEY, J., The future of data analysis. *Annals of Mathematical Statistics*, 1962, **33**, 1–67.

CHAPTER 20

ESTIMATION OF AN ABILITY

20.1 Introduction

In this chapter we shall systematically develop some general methods of assessing the usefulness of test models for estimating a subject's ability θ. We shall then show that these methods provide general guidance and specific working techniques for selecting items and for designing and constructing tests for specified purposes.

We shall develop these methods in terms of the normal approximation to the distribution of scoring formulas $x = x(\mathbf{v})$, and in terms of point and confidence limit estimators $\theta^*(x, \alpha)$ of θ, which we introduced in Sections 17.4 and 17.7. There we saw that the precision properties of estimators based on a given scoring formula are usefully represented by (1) the variance of the scoring formula $\sigma^2[x(\mathbf{V}), \theta]$, and (2) the derivative $\partial \mathscr{E}[x(\mathbf{V}) \mid \theta]/\partial\theta$, which specifies how the mean of the scoring formula depends on θ.

In particular, we saw that these precision properties are summarized in the *information function of a given scoring formula*

$$I[\theta, x(\mathbf{v})] = \frac{1}{\sigma^2[x(\mathbf{V}), \theta]} \left\{ \frac{\partial}{\partial\theta} \mathscr{E}[x(\mathbf{V}) \mid \theta] \right\}^2, \tag{20.1.1}$$

where $x = x(\mathbf{v})$ denotes any given test scoring formula based on any given test model. The model may be represented by its pdf's Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ or just the cdf's $F(x \mid \theta)$ of the given score. For brevity, we shall write $I[\theta, x(\mathbf{v})] = I(\theta, x)$.

20.2 Some Algebra of Information Functions

If $x(\mathbf{v})$ has any weighted-sum form

$$x(\mathbf{v}) = \sum_{g=1}^{n} w_g u_g, \tag{20.2.1}$$

where the w_g are any positive numbers, then we may substitute (17.7.2) and (17.7.3) in (20.1.1) and obtain

$$I(\theta, x) = \left[\sum_{g=1}^{n} w_g^2 P_g(\theta) Q_g(\theta) \right]^{-1} \left[\sum_{g=1}^{n} w_g P_g'(\theta) \right]^2, \tag{20.2.2}$$

where

$$P'_g(\theta) = \frac{\partial}{\partial\theta} P_g(\theta).$$

As a case of the score formula $x(\mathbf{v})$ we may take a single term $w_g u_g$ of such a score formula or a single item response u_g, either of which gives the *item information function*

$$I(\theta, u_g) \equiv I(\theta, w_g u_g) = P'_g(\theta)^2 / P_g(\theta) Q_g(\theta). \tag{20.2.3}$$

Now we may easily verify (by application of the Cauchy inequality) that for any given numbers C_g, $D_g (D_g > 0)$, we have

$$\left(\sum_{g=1}^{n} w_g C_g\right)^2 \leqq \sum_{g=1}^{n} (C_g/\sqrt{D_g})^2 \sum_{g=1}^{n} (w_g\sqrt{D_g})^2, \tag{20.2.4}$$

with equality if and only if $w_g = AC_g/D_g$, $g = 1, \ldots, n$, where A may be any nonzero number. Setting

$$C_g = P'_g(\theta) \qquad \text{and} \qquad D_g = P_g(\theta)Q_g(\theta) \text{ gives}$$

$$I(\theta, x) \leq \sum_{g=1}^{n} I(\theta, u_g), \tag{20.2.5}$$

with equality if and only if

$$w_g = P'_g(\theta)/P_g(\theta)Q_g(\theta) = w_g(\theta), \qquad g = 1, \ldots, n, \tag{20.2.6}$$

except for the usual allowed arbitrary positive constant factor. That is, equality obtains if and only if the w_g are locally best weights at θ. We noted in Section 19.3 that locally best weights independent of θ exist only in logistic test models. In general, then,

$$I(\theta, x) < \sum_{g=1}^{n} I(\theta, u_g),$$

except at those values of θ, if any, for which the given weights are locally best; and equality obtains uniformly in θ only if the model is logistic (or consists of equivalent items) and if $x = \sum_{g=1}^{n} a_g u_g$. We shall call the right-hand member of (20.2.5),

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g) \equiv \sum_{g=1}^{n} [P'_g(\theta)]^2 / P_g(\theta)Q_g(\theta), \tag{20.2.7}$$

the *information function of a test.*

We note, therefore, that $I(\theta)$ is determined by the test model, since it is only the sum of the information functions of its items, and that it is not dependent on

any choice of a score formula $x(\mathbf{v})$. Because of these facts and because of the relation (20.2.5), it constitutes an upper bound on each and all of the information functions $I(\theta, x)$ that may be obtained by the various possible choices of test score formulas of the weighted sum form.

When no score formula of the weighted sum form gives equality for all θ in (20.2.5), the precision of estimation represented by such equality is nevertheless usually obtainable by use of statistics of more widely applicable forms such as maximum likelihood estimators, as we shall show in the next section. Also, since the formula (20.2.7) exhibits a basic general additivity property of the information structures of tests in relation to their items, it is a basis for solving a number of problems of appraising and designing of tests, test items, and score formulas and estimators, as we shall show in Sections 20.4 through 20.6.

20.3 More General Methods of Estimation: Maximum Likelihood

The relation

$$I(\theta, x) = I(\theta)$$

never holds, uniformly in θ, for linear $x(\mathbf{v})$ except in logistic or equivalent-item test models, as we have indicated in the preceding section. For test models of other forms, the precision of estimation represented by this relation is obtainable in many cases of interest by use of estimators based on certain *nonlinear* score formulas $x(\mathbf{v})$. Since each point estimator $\theta^* = \theta^*(x) = \theta^*[x(\mathbf{v})]$ based on a statistic $x = x(\mathbf{v})$ is itself a real-valued statistic $\theta^* = \theta^*[\mathbf{v}]$, it is appropriate and sometimes convenient to express such precision properties of a given estimator θ^* by writing

$$I(\theta, \theta^*[\mathbf{v}]) \equiv I(\theta, \theta^*) = I(\theta), \tag{20.3.1}$$

uniformly in θ. [Here $I(\theta, \theta^*)$ is calculated using the asymptotic mean and variance of θ^*, as discussed in Section 17.7.] In many of these cases we shall see that it is convenient to define and use such estimators without reference to any other intermediary statistic $x(\mathbf{v})$.

Estimators θ^* satisfying (20.3.1) are called *asymptotically efficient*, since the right-hand member of (20.3.1) can be shown to be an upper bound for the information function of any test no matter what estimator is used.* For our purposes, the most convenient of these estimators are the *maximum likelihood estimators*, customarily denoted by $\hat{\theta} = \hat{\theta}(\mathbf{v})$. When $\mathbf{v}$ is fixed at an observed response pattern, the function Prob $(\mathbf{V} = \mathbf{v} \mid \theta)$ of θ is called the *likelihood function*. If, as occurs in most of the models discussed above, this function has, for each $\mathbf{v}$, a maximum that is attained at a unique value of θ, the *maximum likelihood estimate* $\hat{\theta} = \hat{\theta}(\mathbf{v})$ is defined to have this value.

---

* Exceptions, of more theoretical than practical interest, are the *superefficient estimators*. See Kendall and Stuart, 1961, p. 44.

For both theoretical and practical purposes, it is often convenient to avoid writing an explicit formula for $\hat{\theta}(\mathbf{v})$, which is in general nonlinear. Provided that the derivative

$$\frac{\partial}{\partial\theta}\log \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)$$

exists for each $\mathbf{v}$ and is a decreasing function of θ that assumes the value 0, the *likelihood equation*

$$\frac{\partial}{\partial\theta}\log \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) = 0 \tag{20.3.2}$$

defines implicitly the maximum likelihood estimator $\hat{\theta}(\mathbf{v})$; for each observed $\mathbf{v}$, $\hat{\theta} = \hat{\theta}(\mathbf{v})$ can be calculated numerically as the solution of this equation.

Some of the efficiency properties of maximum likelihood estimators can be proved in exact and elementary terms (see Birnbaum, 1961a, pp. 122–127): Let

$$\begin{aligned}\frac{\partial}{\partial\theta}\log \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta) &\equiv \frac{\partial}{\partial\theta}\log \prod_{g=1}^{n}\left[\frac{P_g(\theta)}{Q_g(\theta)}\right]^{u_g}\prod_{g=1}^{n} Q_g(\theta)\\ &= \sum_{g=1}^{n} w_g(\theta)u_g - R(\theta)\\ &= x(\mathbf{v}, \theta) - R(\theta),\end{aligned} \tag{20.3.3}$$

where

$$w_g(\theta) = \frac{\partial}{\partial\theta}\log\frac{P_g(\theta)}{Q_g(\theta)} \qquad \text{and} \qquad R(\theta) = -\frac{\partial}{\partial\theta}\log\prod_{g=1}^{n} Q_g(\theta),$$

$w_g(\theta)$ being a locally best weight (see Section 19.3) and $R(\theta)$ being a nonnegative quantity. We note that $x(\mathbf{v}, \theta)$ is not a statistic, because of its dependence on θ; but if θ is fixed at any chosen value θ', then $x(\mathbf{v}, \theta')$ is a statistic, and in particular it is the locally best score formula at θ'.

One locally best classification rule for abilities in the neighborhood of θ' has the form: Classify as high just when $x(\mathbf{v}, \theta') > R(\theta')$. Now if $\mathbf{v}'$ is any response pattern that is classified as high by this rule, we have $x(\mathbf{v}', \theta') > R(\theta')$. Under the conditions mentioned above, $x(\mathbf{v}', \theta)$ is decreasing in θ, and hence $\hat{\theta}(\mathbf{v}') > \theta'$. Similarly, if $\mathbf{v}''$ is any response pattern that is classified as low by the rule described, we have $\hat{\theta}(\mathbf{v}'') \leq \theta'$. Thus we see that if we compare any maximum likelihood estimate $\hat{\theta}(\mathbf{v})$ with any specified ability level θ' of interest, and if we take $\hat{\theta}(\mathbf{v}) > \theta'$ as evidence that the true ability value exceeds θ' and in similar fashion take $\hat{\theta}(\mathbf{v}) \leq \theta'$ as evidence that the true ability value does not exceed θ', then we are interpreting maximum likelihood estimates in ways that correspond in formal detail with certain locally best classification rules.

In cases in which the locally best score has approximately a normal distribution at θ' (see Section 17.7), the probability Prob $[x(\mathbf{V}, \theta') > R(\theta') \mid \theta']$, which uniquely characterizes one locally best rule, will be very near $\frac{1}{2}$. To

prove this, we note that

$$\begin{aligned}
\mathscr{E}[x(\mathbf{V}, \theta') \mid \theta'] &= \sum_{\mathbf{v}} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta')x(\mathbf{v}, \theta') \\
&= \sum_{\mathbf{v}} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta')\left[\frac{\partial}{\partial\theta}\log \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)\Big|_{\theta=\theta'} + R(\theta')\right] \\
&= \sum_{\mathbf{v}} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta') \\
&\quad \times \left[\frac{1}{\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta')}\frac{\partial}{\partial\theta}\text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)\Big|_{\theta=\theta'}\right] + R(\theta') \\
&= \frac{\partial}{\partial\theta}\left[\sum_{\mathbf{v}} \text{Prob}\,(\mathbf{V} = \mathbf{v} \mid \theta)\right]\Big|_{\theta=\theta'} + R(\theta') \\
&= \frac{\partial}{\partial\theta}1 + R(\theta') = R(\theta'), \qquad (20.3.4)
\end{aligned}$$

and hence

$$\text{Prob}\,[x(\mathbf{V}, \theta') \leq R(\theta') \mid \theta'] \doteq \Phi\left[\frac{R(\theta') - \mathscr{E}[x(\mathbf{V}, \theta') \mid \theta']}{\sigma[x(\mathbf{V}, \theta') \mid \theta']}\right] = \Phi(0) = \frac{1}{2}. \qquad (20.3.5)$$

Combining this approximate value with the observation above, that $x(\mathbf{v}, \theta') > R(\theta')$ for each θ' if and only if $\hat{\theta}(\mathbf{v}) > \theta'$, we see that under the conditions mentioned, the maximum likelihood estimator is approximately median-unbiased. (The latter property was defined in Section 17.4.)

It can further be shown (Cramér, 1946, p. 500) that *a maximum likelihood estimator has approximately (asymptotically) the normal distribution with mean θ, the true ability value, and variance $1/I(\theta)$, under conditions satisfied by most of the models under discussion.* Calculating $I(\theta, \hat{\theta})$ with these asymptotic values, we find that $I(\theta, \hat{\theta}) = I(\theta)$, and therefore *the maximum likelihood estimator $\hat{\theta}$ is asymptotically efficient.*

In those special cases in which $I(\theta)$ is not dependent on the unknown true value of θ, its value I can be computed and used in formulas for such statistics as

$$\hat{\theta} + \Phi^{-1}(1 - \alpha)/\sqrt{I(\theta)}, \qquad (20.3.6)$$

which represents a *maximum likelihood confidence limit estimator* with approximate (asymptotic) confidence coefficient $1 - \alpha$ (or α, if $1 - \alpha < \frac{1}{2}$). Such an estimator shares the asymptotic efficiency property of $\hat{\theta}$ itself. Fortunately, although $I(\theta)$ varies with θ in most cases of interest here, it can be shown (Kendall and Stuart, 1961; Wald, 1942) that replacing the unknown θ in $I(\theta)$ by its estimate $\hat{\theta}$ and substituting in (20.3.6) gives a *maximum likelihood confidence limit estimator*

$$\hat{\theta}(\mathbf{v}, 1 - \alpha) \equiv \hat{\theta} + \Phi^{-1}(1 - \alpha)/\sqrt{I(\hat{\theta})}, \qquad (20.3.7)$$

which also shares the asymptotic efficiency property of $\hat{\theta}$ under regularity conditions satisfied by most of the models considered here.

Examples of maximum likelihood estimators of an ability

1. Logistic test model

$$x(\mathbf{v}, \theta) = \sum_{g=1}^{n} w_g(\theta) u_g, \tag{20.3.8}$$

where

$$w_g(\theta) = \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} = \frac{\partial}{\partial \theta} \log \frac{\Psi[DL_g(\theta)]}{\Psi[-DL_g(\theta)]}$$

$$= \frac{\partial}{\partial \theta} \log e^{DL_g(\theta)} = \frac{\partial}{\partial \theta} DL_g(\theta) = Da_g,$$

and

$$R(\theta) = -\sum_{g=1}^{n} \frac{\partial}{\partial \theta} \log Q_g(\theta),$$

in which

$$-\frac{\partial}{\partial \theta} \log Q_g(\theta) = -\frac{\partial}{\partial \theta} \log [1 + e^{DL_g(\theta)}]^{-1} = Da_g \Psi[DL_g(\theta)] = Da_g P_g(\theta).$$

Thus the likelihood equation is

$$x(\mathbf{v}, \hat{\theta}) - R(\hat{\theta}) = D \sum_{g=1}^{n} a_g u_g - D \sum_{g=1}^{n} a_g P_g(\hat{\theta}) = 0, \tag{20.3.9}$$

or

$$x(\mathbf{v}) \equiv \sum_{g=1}^{n} a_g u_g = \sum_{g=1}^{n} a_g P_g(\hat{\theta}) \equiv E\left(\sum_{g=1}^{n} a_g U_g \mid \hat{\theta}\right),$$

or

$$\sum_{g=1}^{n} a_g \Psi[DL_g(\hat{\theta})] = \sum_{g=1}^{n} a_g u_g. \tag{20.3.10}$$

The right-hand member of the equation (20.3.10) is simply a numerical value of the familiar logistic test score, and the left-hand member is a strictly increasing function of $\hat{\theta}$. The reader should note that $\hat{\theta}$ is a one-to-one function of the sufficient statistic $x(\mathbf{v})$. No convenient explicit formula for $\hat{\theta}$ is available for the general logistic test model.

In the special case of equivalent items, the equation becomes

$$P(\hat{\theta}) = \overline{u}, \qquad \text{where} \qquad \overline{u} = \frac{1}{n} \sum_{g=1}^{n} u_g, \tag{20.3.11}$$

and where $P(\theta)$ is the common item characteristic curve. This case admits the explicit solution for the maximum likelihood estimator

$$\hat{\theta} = P^{-1}(\overline{u}) = b + \frac{\Psi^{-1}(\overline{u})}{Da} = b + \frac{1}{Da} \log\left(\frac{\overline{u}}{1 - \overline{u}}\right). \tag{20.3.12}$$

The explicit formula $\hat{\theta} = P^{-1}(\overline{u})$ also holds for other cases of tests composed

of equivalent items for which the common item characteristic curves have any form $P(\theta)$ increasing strictly and continuously from zero to one.

The likelihood equation for a general logistic test model can be solved conveniently in the form (20.3.10) by evaluating the left-hand member of the equation (which increases strictly with θ) at trial values of θ, using available tables of $\Psi(t)$ (see Berkson, 1957, pp. 33–34). Alternatively, if maximum likelihood estimation is to be used repeatedly with a given test model, one can prepare a table or graph of the left-hand member of the likelihood function, from which one can read, as accurately as he desires, the maximum likelihood estimate $\hat{\theta} = \hat{\theta}(x)$ corresponding to each possible value of the score $x = \sum_{g=1}^{n} a_g u_g$. Such a graph corresponds approximately to the contour $F(x \mid \theta) = 0.5$ in Fig. 17.4.2, which was used there to determine median-unbiased estimators; this corresponds to the above-mentioned approximate median unbiasedness that holds for maximum likelihood estimators in many cases of interest.

To determine maximum likelihood confidence limits as given by (20.3.7), we must compute

$$I(\hat{\theta}) = D^2 \sum_{g=1}^{n} a_g^2 \psi[DL_g(\hat{\theta})].$$

This computation can be facilitated by use of available tables of $\psi(t)$ (Berkson, 1957). Again, for convenience in repeated use, one may prepare a table or graph of the function $I(\hat{\theta})^{-1/2}$, the asymptotic estimate of standard error that appears with the maximum likelihood estimate $\hat{\theta}$ in the formula (20.3.7) for $\hat{\theta}(\mathrm{v}, 1 - \alpha)$.

2. Normal ogive model. Lord (1953, pp. 60–63) has treated this case in some detail. When specialized to the present case of known item parameters, Lord's treatment of maximum likelihood estimation of ability is analogous to the following treatment of the three-parameter logistic model.

3. Three-parameter logistic model. The general features of maximum likelihood estimation methods here closely resemble those of the special case in which all guessing probabilities are zero, that is, the case of the general logistic model discussed above. The locally best weights are

$$w_g(\theta) \equiv \frac{\partial}{\partial \theta} \log \frac{P_g(\theta)}{Q_g(\theta)} = Da_g \Psi[DL_g(\theta) - \log c_g],$$

and the likelihood equation can be written as

$$\sum_{g=1}^{n} a_g \Psi[DL_g(\hat{\theta})] = \sum_{g=1}^{n} a_g \frac{w_g(\hat{\theta})}{D} u_g.$$

We observe that the left-hand member here is identical to the left-hand member of (20.3.10), the equation for the two-parameter logistic model. The right-hand member varies with $\hat{\theta}$. For a fixed response pattern, $\hat{\theta}$ may be computed numerically by trials, with use of tables of $\Psi(t)$.

20.4 The Information Functions of Various Test Items

The information functions of items may be regarded as building blocks from which the information function of a test is constructed. This is implied by the basic general additive relation represented in the defining equation (20.2.7):

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g) = \sum_{g=1}^{n} [P'_g(\theta)]^2 / P_g(\theta) Q_g(\theta). \tag{20.4.1}$$

In this section we shall prepare ourselves to use this relation systematically by analyzing in detail the contributions of individual items to a test information function. For items of several of the test models considered above, we shall determine the item information function $I(\theta, u_g)$. We shall also determine three parameters of that function which describe aspects of the item's contribution to the test information function:

1. The maximum of $I(\theta, u_g)$,

$$M_g = \max_{\theta} I(\theta, u_g), \tag{20.4.2}$$

will sometimes be of interest, since at each θ we have

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g) \leq \sum_{g=1}^{n} M_g. \tag{20.4.3}$$

2. We shall denote the value of θ at which the maximum is attained by θ_g, when this value is uniquely defined.
3. The area under the curve $I(\theta, u_g)$,

$$A_g = \int_{-\infty}^{\infty} I(\theta, u_g)\, d\theta, \tag{20.4.4}$$

will also be of some interest, since

$$\int_{-\infty}^{\infty} I(\theta)\, d\theta = \sum_{g=1}^{n} A_g. \tag{20.4.5}$$

We shall now determine these parameters for specific test models.

1. Logistic. For

$$P_g(\theta) = \Psi[DL_g(\theta)] = [1 + \exp D(a_g\theta - b_g)]^{-1},$$

we have

$$P'_g(\theta) = Da_g\psi[DL_g(\theta)] \quad \text{and} \quad P_g(\theta)Q_g(\theta) = \psi[DL_g(\theta)]. \tag{20.4.6}$$

Thus the information function of a test item in the logistic model is

$$I(\theta, u_g) = P'_g(\theta)^2 / P_g(\theta) Q_g(\theta) = D^2 a_g^2 \psi[DL_g(\theta)]. \tag{20.4.7}$$

Since the maximum of $\psi[DL_g(\theta)]$ is $\psi(0) = \Psi(0)[1 - \Psi(0)] = (\frac{1}{2})^2 = \frac{1}{4}$, the maximum of $I(\theta, u_g)$ is

$$M_g = \tfrac{1}{4}D^2a_g^2 = \tfrac{1}{4}a_g^2(1.7)^2 = 0.721a_g^2; \tag{20.4.8}$$

it is attained at $\theta_g = b_g$. The area under $I(\theta, u_g)$ is

$$A_g = \int_{-\infty}^{\infty} I(\theta, u_g)\, d\theta = Da_g \int_{-\infty}^{\infty} Da_g\psi[DL_g(\theta)]\, d\theta = Da_g = 1.7a_g. \tag{20.4.9}$$

Figure 20.4.1 shows information curves for logistic items with difficulties b_g all zero, and $a_g = g$, $g = 1, 2, 3$. Values $a_g > 1$ are considered rare in practice; typically, therefore, the area under an item information curve is spread more or less smoothly over at least several units on the θ scale rather than concentrated over a small interval of θ values (as would be theoretically optimal for some of the classification problems discussed in the previous chapter).

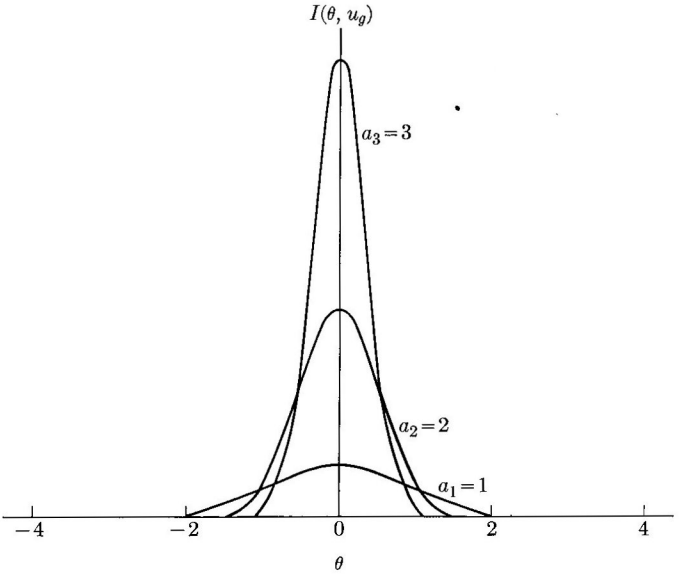

FIG. 20.4.1. Item characteristic curves of the logistic items with $a_1 = 1$, $a_2 = 2$, $a_3 = 3$, and $b_g = 0$, $g = 1, 2, 3$.

2. Normal ogive. We have

$$\begin{aligned} I(\theta, u_g) &= P_g'(\theta)^2/P_g(\theta)Q_g(\theta) \\ &= a_g^2\varphi[L_g(\theta)]^2/\Phi[L_g(\theta)]\Phi[-L_g(\theta)] \end{aligned} \tag{20.4.10}$$

or

$$I(\theta, u_g) = a_g^2J(s), \tag{20.4.11}$$

where

$$J(s) = \varphi(s)^2/\Phi(s)\Phi(-s), \qquad \text{and} \qquad s = L_g(\theta). \tag{20.4.12}$$

The maximum of $I(\theta, u_g)$ is attained when $s = 0$, that is, at $\theta_g = b_g$. The maximum is

$$M_g = a_g^2/(\sqrt{2\pi})^2(\tfrac{1}{2})^2 = 2a_g^2/\pi.$$

The area A_g could be determined by numerical integration; it is not used below.

3. Three-parameter logistic. If we take

$$P_g(\theta) = c_g + (1 - c_g)\Psi[DL_g(\theta)] \tag{20.4.13}$$

(this includes the logistic case above as the special extreme case in which $c_g = 0$), we have

$$P'_g(\theta) = (1 - c_g)Da_g\psi[DL_g(\theta)] \tag{20.4.14}$$

and

$$P_g(\theta)Q_g(\theta) = (1 - c_g)\{\psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)]^2\}. \tag{20.4.15}$$

Hence the information function of an item in the logistic test model with guessing probabilities is

$$I(\theta, u_g) = (1 - c_g)D^2a_g^2\psi^2[DL_g(\theta)]/\{\psi[DL_g(\theta)] + c_g\Psi[-DL_g(\theta)]^2\}. \tag{20.4.16}$$

An alternative form is obtained from (20.4.16) by substituting the relation

$$\begin{aligned}\frac{\psi(t) + c\Psi(-t)^2}{\psi(t)} &= 1 + c\frac{1}{(1 + e^t)}\frac{(1 + e^t)}{e^t} \\ &= 1 + \frac{c}{e^t} = \frac{(e^t/c) + 1}{(e^t/c)} = \frac{e^{t-\log c} + 1}{e^{t-\log c}} \\ &= 1/\Psi(t - \log c).\end{aligned} \tag{20.4.17}$$

This results in

$$I(\theta, u_g) = \{D^2a_g^2(1 - c_g)\psi[DL_g(\theta)]\}\{\Psi[DL_g(\theta) - \log c_g]\}. \tag{20.4.18}$$

In this form, the information function is expressed as a product in which the first factor is the information function of a corresponding hypothetical logistic item with the same parameters, except that $c_g = 0$; this hypothetical item's characteristic curve is thus the probability of a correct response without guessing in the hypothetical interpretation described in Section 17.3. According to this interpretation, the factor $(1 - c_g)$ decreases the factor described by exactly the probability c_g that a subject of any ability who cannot answer correctly without guessing will answer correctly by guessing. The final factor is a cumulative logistic distribution function whose median is $b_g + (\log c_g)/Da_g$, which lies below b_g by an amount proportional to $\log(1/c_g)$; if c_g is very small, this factor is near unity when θ is not far below b_g. If, for example, $c_g = 0.2$, the median is

$$b_g - \frac{(\log_e 5)}{Da_g} = b_g - \frac{1.61}{1.7a_g} = b_g - \frac{0.95}{a_g}. \tag{20.4.19}$$

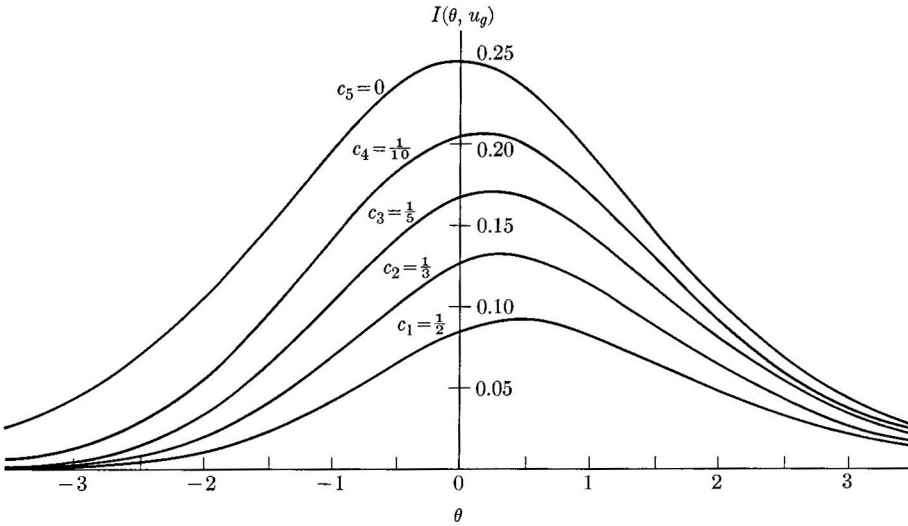

FIG. 20.4.2. Information curves of logistic items with guessing probabilities, with $a_g = 0.588$, $b_g = 0.0$, $g = 1, \ldots, 5$, and $c_1 = \frac{1}{2}$, $c_2 = \frac{1}{3}$, $c_3 = \frac{1}{5}$, $c_4 = \frac{1}{10}$, and $c_5 = 0$.

Then if $a_g = 1/1.7 = 0.588$, $b_g = 0$, and $c_g = 0.2$, we may say that this median is -1.61.

Another useful form is

$$I(\theta, u_g) = D^2 a_g^2 \psi[DL_g(\theta)] - D^2 a_g^2 P_g(\theta) \psi[DL_g(\theta) - \log c_g]. \quad (20.4.20)$$

Here the first term on the right-hand side is the information function of the hypothetical logistic item referred to, and the second term represents in an additive form the information hypothetically lost because of guessing.

Figure 20.4.2 shows several such information curves for items with $b_g = 0$ and $a_g = 1/D = 1/1.7 = 0.588$, with c_g taking the values $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{5}$, and $\frac{1}{10}$, respectively. The limiting case $c_g = 0$ is included for comparison.

Later in this section we shall show that the value θ_g of θ that maximizes $I(\theta, u_g)$ is

$$\theta_g = b_g + \frac{1}{Da_g} \log \frac{1 + (1 + 8c_g)^{1/2}}{2}. \quad (20.4.21)$$

Since $0 < c_g < 1$, we have

$$0 < \log \frac{1 + (1 + 8c_g)^{1/2}}{2} < \log_e 2 = 0.69.$$

Figure 20.4.2 illustrates the increase of θ_g with c_g for items with $a_g = 1/D$ and $b_g = 0$.

If any number n of items were available with given common values of a_g and c_g and with the values b_g, $g = 1, \ldots, n$, subject to our specification, then, according to (20.4.21), we should specify the common value

$$b_g = \theta' - \frac{1}{Da_g} \log \frac{1 + (1 + 8c_g)^{1/2}}{2}, \qquad g = 1, \ldots, n, \tag{20.4.22}$$

to maximize

$$I(\theta) = \sum_{i=1}^{n} I(\theta, u_g)$$

at a given ability level θ'. Conclusions very similar to these, quantitatively as well as qualitatively, have been obtained for analogous efficiency problems in the normal ogive test model with guessing probabilities, under the assumption that abilities have a given probability distribution (Lord, 1953, pp. 67–69, and references therein).

To derive the formula (20.4.21) for θ_g, we write $t = DL_g(\theta) = Da_g(\theta - b_g)$ and maximize $I(\theta, u_g)$ with respect to t. From (20.4.18), we find that

$$\begin{aligned}
\frac{\partial}{\partial t} \log I(\theta, u_g) &= \frac{\partial}{\partial t} [\log \psi(t) + \log \Psi(t - \log c_g)] \\
&= 2\Psi(-t) - 1 + \Psi(-t + \log c_g) \\
&= \frac{2}{1 + e^t} - 1 + \frac{1}{1 + e^t/c_g} = \frac{1 - e^t}{1 + e^t} + \frac{c_g}{c_g + e^t} \\
&= (2c_g + e^t - e^{2t})/(c_g + e^t)(1 + e^t).
\end{aligned} \tag{20.4.23}$$

Setting this derivative equal to zero and solving for t yields

$$t \equiv Da_g(\theta_g - b_g) = \log (\tfrac{1}{2} + \tfrac{1}{2}\sqrt{1 + 8c_g}), \tag{20.4.24}$$

which gives the stated formula for θ_g.

For the area A_g under an item information curve, we have

$$\begin{aligned}
A_g &= \int_{-\infty}^{\infty} I(\theta, u_g)\, d\theta \\
&= D^2 a_g^2 (1 - c_g) \int_{-\infty}^{\infty} \Psi[DL_g(\theta)] \Psi[DL_g(\theta) - \log c_g]\, d\theta.
\end{aligned} \tag{20.4.25}$$

After making the substitution $r = 1 + \exp DL_g(\theta)$ and integrating by the method of partial fractions, we find that

$$A_g = Da_g \frac{c_g \log c_g + 1 - c_g}{(1 - c_g)}. \tag{20.4.26}$$

As $c_g \to 0$, the right-hand member increases to Da_g, the exact value (Eq. 20.4.9) of A_g found for two-parameter logistic items.

20.5 The Information Functions of Various Tests

In the present section we shall describe some examples of test information curves as they are related to item information curves by the additive definition (20.2.7):

$$I(\theta) = \sum_{g=1}^{n} I(\theta, u_g).$$

The examples will also illustrate the discussion of problems of test design and item selection in the following section.

1. Tests with equivalent items. If $P_1(\theta)$ is the common ICC of the n items constituting a test, and if $I(\theta, u_1)$ is the common item information function determined by $P_1(\theta)$, then the information function of the test is

$$I(\theta) = nI(\theta, u_1). \tag{20.5.1}$$

If M_1 denotes the maximum of $I(\theta, u_1)$, and that maximum is attained at $\theta = \theta_1$, then the maximum of $I(\theta)$ is nM_1, and is also attained at $\theta = \theta_1$. If A_1 is the area under $I(\theta, u_1)$, then the area under $I(\theta)$ is nA_1. Each detail of an item information function thus determines in a simple way a corresponding detail of the information function of a test consisting of any number n of such equivalent items.

2. Tests with nonequivalent items. A hypothetical test consisting of only several logistic items, with a_g-values appreciably above the range encountered in practice, and with unequal b_g-values, would have an information function very near zero except for a very high peak near each b_g-value. This configuration reflects the function's high discriminating power in the neighborhood of each b_g-value but otherwise low precision. We discussed such a hypothetical test in Section 16.5, where we noted that items with extremely high a_g-values, if available in practice, would be ideal for classification into ordered intervals of ability levels, but that they would be of limited value for estimation of ability with good precision over an appreciably wide range of θ-values. In a somewhat different theoretical context (where some probability distribution of abilities is assumed), there is a logical possibility that a sufficiently extreme increase of a_g-values, while other parameters remain fixed, could result in lowering the precision of estimation obtainable with a given test. This possibility has been recognized and discussed by Tucker (1946), Loevinger (1954, as "The Attenuation Paradox in Test Theory"), Solomon (1956), and Sitgreaves (1961). A number of authors have suggested that this "paradox" can be resolved or avoided by use of items with suitably spaced item difficulty parameters b_g.

Clearly, if each a_g is extremely high, then one has virtually error-free discrimination between ability levels above and below each b_g-value represented in the test, and virtually no other information for discrimination or estimation. However, in the cases encountered in practice, a_g-values exceeding unity are rare; the total area $1.7a_g$ under the item information function typically is spread rather smoothly over at least several units on the θ scale, with only

moderate concentration around b_g. Thus the greatest concentration practically attainable of the area under a test information curve is obtained by taking items with some common difficulty level, along with a_g-values as high as possible; but this gives at best only the same limited degree of concentration found in the information curves of the individual items with the highest a_g-values. *Thus test information curves in practice tend to be less concentrated than item information curves, by an amount that depends primarily on the variation among item difficulty parameters.*

A rough but useful concept here is that the area under the information curve tends to be a smoothed version of the histogram formed by adding, for each item in a logistic test model, a square of area Da_g, centered at $\theta = b_g$. Somewhat refined examples of the use of this concept are given next.

3. Uniformly distributed item difficulties. Because there is some tendency in practice for discrimination parameters a_g in normal ogive or logistic models to have a limited range, as mentioned above, while item difficulty parameters b_g tend to vary appreciably, it is of interest to consider test models with common a_g-values but with unequal b_g-values distributed in certain regular patterns. Such oversimplified, schematized versions of cases encountered in practice can provide useful insights, approximations, and guidance.

Consider a logistic test model with n items having discrimination parameters with the common value a_1, and with difficulty parameters b_g uniformly spaced, so that, for some $\epsilon > 0$,

$$b_2 = b_1 + \epsilon, \quad b_3 = b_1 + 2\epsilon, \quad \ldots, \quad b_n = b_1 + (n-1)\epsilon.$$

If n is not small and a_1 is not large, and if the range of difficulty parameters $(n-1)\epsilon$ is not narrow, then the continuous uniform probability density function

$$q(b) = \begin{cases} 1/n\epsilon & \text{for } b_1 - \epsilon/2 \leq b \leq b_n + \epsilon/2, \\ 0 & \text{for other values of } b, \end{cases} \tag{20.5.2}$$

will give a rough approximation to the formal discrete distribution function of the b_g-values, which will, in turn, give useful approximation formulas for $I(\theta)$. We have

$$\begin{aligned} I(\theta) &= D^2 a_1^2 \sum_{g=1}^{n} \psi[Da_1(\theta - b_g)] \doteq D^2 a_1^2 n \int_{b_1-\epsilon/2}^{b_n+\epsilon/2} \psi[Da_1(\theta - b_g)] q(b)\, db \\ &= \frac{Da_1}{\epsilon} \int_{b_1-\epsilon/2}^{b_n+\epsilon/2} Da_1 \psi[Da_1(\theta - b)]\, db \\ &= \frac{Da_1}{\epsilon} \left\{ \Psi\left[Da_1 \left(b_n + \frac{\epsilon}{2} - \theta \right) \right] - \Psi\left[Da_1 \left(b_1 - \frac{\epsilon}{2} - \theta \right) \right] \right\} \\ &\doteq \frac{1.7a_1}{\epsilon} \left\{ \Phi\left[a_1 \left(b_n + \frac{\epsilon}{2} - \theta \right) \right] - \Phi\left[a_1 \left(b_1 - \frac{\epsilon}{2} - \theta \right) \right] \right\}. \end{aligned} \tag{20.5.3}$$

Thus, for θ well within the range of b_g-values, under the conditions indicated, the last factor will be near unity. Therefore

$$I(\theta) \doteq \frac{1.7a_1}{\epsilon}, \qquad \text{for } b_1 + \frac{2}{a_1} < \theta < b_n - \frac{2}{a_1}; \tag{20.5.4}$$

that is, $I(\theta)$ will have approximately the constant value indicated, which is proportional to the number $(1/\epsilon)$ of items located (by their difficulty levels b_g) on each unit of the θ scale and to the area $1.7a_1$ under an individual item information curve.

4. Normally distributed item difficulties. Evidently there is some tendency toward limited concentration of the difficulty parameters b_g in cases encountered in practice. Let us consider a schematized example that represents this feature. Suppose that a test's n items have the identical parameters $a_g \equiv a_1$, $g = 1, \ldots, n$, and that the item difficulty parameters b_g, $g = 1, \ldots, n$, are spread approximately in the form of a normal distribution with mean and variance

$$\bar{b} = \frac{1}{n}\sum_{g=1}^{n} b_g \qquad \text{and} \qquad \sigma_b^2 = \frac{1}{n}\sum_{g=1}^{n} (b_g - \bar{b})^2,$$

respectively. This normal distribution is represented by the pdf

$$q(b) = \frac{1}{\sigma_b}\varphi\left(\frac{b - \bar{b}}{\sigma_b}\right). \tag{20.5.5}$$

Both Lawley (1943) and Lord (1952, 1953) have considered this case for the normal ogive model. For the logistic model, we have

$$\begin{aligned} I(\theta) &= D^2a_1^2 \sum_{g=1}^{n} \psi[Da_1(\theta - b_g)] \doteq D^2a_1^2 n \int_{-\infty}^{\infty} \psi[Da_1(\theta - b)]q(b)\, db \\ &= D^2a_1^2 \frac{n}{\sigma_b}\int_{-\infty}^{\infty} \psi[Da_1(\theta - b)]\varphi\left(\frac{b - \bar{b}}{\sigma_b}\right) db. \end{aligned} \tag{20.5.6}$$

An approximation that plausibly is close here is the one obtained by replacing the logistic by a normal density function:

$$\begin{aligned} I(\theta) &\doteq Da_1 \frac{n}{\sigma_b}\int_{-\infty}^{\infty} a_1\varphi[a_1(\theta - b)]\varphi\left(\frac{b - \bar{b}}{\sigma_b}\right) db \\ &= Da_1^2 \frac{n}{2\pi\sigma_b}\int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}a_1^2(\theta - b)^2\right] \exp\left[-\frac{1}{2\sigma_b^2}(b - \bar{b})^2\right] db \\ &= Da_1 \frac{n}{h}\varphi\left(\frac{\theta - \bar{b}}{h}\right), \qquad \text{where } h^2 = \sigma_b^2 + \frac{1}{a_1^2}. \end{aligned} \tag{20.5.7}$$

That is, when a test is constituted by an appreciable number n of items with common discrimination parameter a_1 and with various difficulty parameters b_g

having a distribution that formally approximates a normal distribution with mean $\bar{b}$ and variance σ_b^2, *the test information curve has approximately the form of a normal pdf*, with mean $\bar{b}$ and variance $h^2 = \sigma_b^2 + 1/a_1^2$, multiplied by the constant nDa_1.

20.6 Problems of Test Design and Item Selection

In models considered in practice, we usually have $30 \leq n \leq 100$. Also, if abilities are scaled so that their distribution is normal ($\mu = 0$, $\sigma^2 = 1$) over the population of interest, then we usually have $a_g < 2$, where a_g is the discrimination parameter of a logistic item with characteristic curve $\Psi[1.7a_g(\theta - b_g)]$. The difficulty parameters b_g usually vary over the range $-3 < b_g < 3$; however, when a_g is close to zero, the b_g may greatly exceed 3 in absolute value. For such models, the approximations described in the preceding section tend to be close. In many cases, the information function of such a model will be approximately equal to those of several other such models, as we shall illustrate.

If two test models have approximately equal information functions $I_1(\theta)$ and $I_2(\theta)$, respectively, then the test models are approximately equivalent, in terms of those statistical precision properties of estimators and classification rules characterized approximately by information functions. It follows that problems of designing a test model that has an approximately prescribed precision, in the sense of a prescribed information function, typically have nonunique solutions; this allows choice among indicated alternative models, based on such considerations as convenience, availability of items of various types, and length of tests.

For individual logistic items, the range indicated above for n and a_g entails the bounds

$$(1.7)(0.2) < A_g < (1.7)(1), \quad \text{or} \quad 0.34 < A_g < 1.7, \tag{20.6.1}$$

on the area A_g under an item information function, and the bounds

$$M_g \equiv \tfrac{1}{4}(1.7)^2 a_g^2 < \tfrac{1}{4}(1.7)^2(1), \quad \text{or} \quad M_g < 0.72, \tag{20.6.2}$$

on the maximum M_g of an item information function. For test information functions, this range entails, as rather wide bounds,

$$(30)(0.34) < \sum_g A_g < 100(1.7), \quad \text{or} \quad 10 < \sum_g A_g < 170, \tag{20.6.3}$$

and

$$I(\theta) < 100(0.72), \quad \text{or} \quad I(\theta) < 72, \quad \text{for all } \theta. \tag{20.6.4}$$

The following examples illustrate the relevance of such bounds and inequalities for problems of design of test models.

1. Consider the requirement that the asymptotic variance of the maximum likelihood estimator of θ, on the basis of a logistic test model, be no greater

than $0.04 = (0.2)^2$ in the neighborhood of $\theta = 1$; that is,

$$1/I(1) \leq 0.04, \qquad \text{or} \qquad I(1) \geq 25.$$

This entails $\sum_{g=1}^{n} M_g \geq 25$; this, in combination with the above bound $M_g < 0.72$, entails $n > 25/0.72$, or $n \geq 35$. The requirement could be met by a logistic model with about $n = 35$ items, provided that all the item difficulty parameters b_g were concentrated very near $\theta = 1$ and all the item discrimination parameters a_g were very nearly as high as unity.

2. Consider the requirement that the precision indicated in (1) be attained with a logistic test model not only at $\theta = 1$, but for $0 < \theta < 3$; that is, $I(\theta) \geq 25$ for $0 < \theta < 3$. This entails

$$\int_0^3 I(\theta)\, d\theta \geq 75 \qquad \text{or} \qquad \sum_{g=1}^{n} A_g \geq 75.$$

This, in combination with the bound $A_g < 1.7$, entails $n > 75/1.7$, or $n \geq 45$. This area requirement could be met by using about 45 items, if items with a_g-values as high as unity were available.

The problem of designing a logistic model whose information curve meets certain given requirements can usefully be regarded as a problem of filling the area under a target information curve by adding contributions of areas $A_g = 1.7a_g$ from respective items, each more or less concentrated about a respective difficulty level b_g. From this standpoint, each of the following groups of logistic items would contribute the same total area toward $\sum_{g=1}^{n} A_g$, namely, 5.1 units of area:

30	items with	$a_g = 0.10$,	5	items with	$a_g = 0.60$,
21	items with	$a_g = 0.14$,	3	items with	$a_g = 1.00$,
15	items with	$a_g = 0.20$,	1	item with	$a_g = 3.00$.
9	items with	$a_g = 0.33$,			

Such a scale for appraising the relative values of items requires the qualification that it may exaggerate the usefulness of items with low a_g values, in the following respect: If $a_g \leq 0.33$, then more than 30% of the area under a logistic item information curve lies outside the interval $-3 < \theta < 3$, to which interest is usually restricted, and so will not contribute to filling the area under a target information curve in problems such as (2) above. Again, if $a_g \leq 0.2$, more than 55% of the area falls outside the interval $-3 < \theta < 3$. On the other hand, for a_g near unity, if b_g satisfies $-1 < b_g < 1$, then less than 5% of the area falls outside the interval $-3 < \theta < 3$.

3. Consider that we specify a target test information function having the form of a given normal pdf with mean μ and standard deviation σ, multiplied by

a positive constant C appreciably larger than unity. We require one or more logistic test models with information curves satisfying

$$I(\theta) \geq \frac{C}{\sigma}\varphi\left(\frac{\theta-\mu}{\sigma}\right),$$

preferably with approximate equality. To fill the area C under the target curve using logistic items with $a_g < 1$, more than $C/(1.7)(1) = C/1.7$ items are required; using items with $a_g < 0.5$, more than $C/(1.7)(0.5) = C/0.85$ items are required.

If the prescribed value of σ is such that $1/\sigma$ is exceeded by the a_g-values of available items, we may take $n \doteq C/Da_1$ items with $a_g = a_1$, $g = 1, \ldots, n$, and with b_g-values varying approximately according to a normal distribution, with mean and variance

$$\bar{b} \equiv \frac{1}{n}\sum_{g=1}^{n} b_g = \mu \qquad \text{and} \qquad \sigma_b^2 \equiv \frac{1}{n}\sum_{g=1}^{n}(b_g - \bar{b})^2 = \sigma^2 - \frac{1}{a_1^2},$$

respectively. Applying the result of example (4) of Section 20.5, we see that this gives a test model with information curve approximately equal to the target. The nonuniqueness of solutions of test design problems mentioned above is illustrated here by the fact that we may take items with any common value $a_g \equiv a_1$, subject to $a_1 > 1/\sigma$.

By taking $n \doteq C$ equivalent logistic items with $a_g = 1/\sigma$ and $b_g = \mu$, $g = 1, \ldots, n$, we would obtain a test with

$$I(\theta) = nDa_1\psi[Da_1(\theta - b_1)] \doteq \frac{C}{\sigma}\varphi\left(\frac{\theta-\mu}{\sigma}\right),$$

where the last approximation is as rough as that of a logistic to a normal density function. The last approximation can be improved somewhat by introducing some variation among the b_g, as discussed above. If the prescribed value σ is such that $1/\sigma$ appreciably exceeds the values a_g of the logistic items available, we see that no close approximation to the target curve is possible; by taking enough items with $b_g = \mu$ to equal the target curve at $\theta = \mu$ so that

$$\sum_{g=1}^{n} Da_g\psi(0) = \frac{D}{4}\sum_{g=1}^{n} a_g \geq \frac{C}{\sigma}\varphi(0) = \frac{C}{\sigma\sqrt{2\pi}},$$

we necessarily obtain a test model whose information curve appreciably exceeds the target for θ not near μ.

4. Consider a requirement of precision expressed in terms of a function $G(\theta)$ which is to be estimated. Let $I[G] \equiv I[G(\theta)]$ denote the information function of any given test, any given increasing differentiable function $G = G(\theta)$

being taken as the unknown parameter of interest. For example, we might represent the quantile-score ranking of an individual with ability θ in a population where abilities have a standard normal distribution by the function $G(\theta) = \Phi(\theta)$.

We have the identity

$$I[G] \equiv \frac{I(\theta)}{[\partial G(\theta)/\partial\theta]^2}.$$

Taking $G(\theta) = \Phi(\theta)$ gives $I[G] = I(\theta)/\varphi(\theta)^2$ or $I(\theta) = I[G]\varphi(\theta)^2$. By the last relation, any requirement of precision of estimation of G, stated in terms of a target information function $I^*[G]$ for $I[G]$, can be restated in terms of a target information function $I^*(\theta)$ for $I(\theta)$, namely, $I^*(\theta) = I^*[G]\varphi(\theta)^2$. This form, with θ as argument, is convenient if item information functions and their properties happen to be available in a form with θ rather than G as argument, as they are in our discussions above.

If we take $I^*[G] \equiv K^2$, where K is any positive constant, the requirement is: The estimation of G with asymptotic variance not to exceed $1/K^2$. This gives

$$I^*(\theta) = K^2\varphi(\theta)^2 = \frac{K^2}{2\pi}(e^{-\theta^2/2})^2 = \frac{K^2}{2\pi}e^{-\theta^2} = \frac{K^2}{2\sqrt{\pi}}\frac{\sqrt{2}}{\sqrt{2\pi}}e^{-(\sqrt{2}\theta)^2/2}$$

$$= \frac{K^2}{2\sqrt{\pi}}\sqrt{2}\varphi(\sqrt{2}\theta).$$

The latter is a case of the above example (3), with $C = K^2/2\sqrt{\pi}$, $\mu = 0$, and $\sigma = 1/\sqrt{2}$; and our discussion of that example applies also to the present one. If $K = 0.02$, the requirement is: The estimation of an examinee's percentile ranking in such a population, with standard error not to exceed 2%. We note again that the techniques of estimation described here, and throughout Chapters 17 through 20, do not depend on assumptions and interpretations in terms of a population distribution of abilities; and that when in fact such assumptions hold with respect to examinees whose abilities are to be estimated, then it is more appropriate and efficient to use estimation methods of the kind introduced by Birnbaum (1967).

20.7 Relative Precisions or Efficiencies of Various Test Designs, Test Score Formulas, and Estimators

Let

$$I_1(\theta, x_1) \equiv I_1[\theta, x_1(\mathbf{v})] \quad \text{and} \quad I_2(\theta, x_2) \equiv I_2[\theta, x_2(\mathbf{v})],$$

respectively, denote the information functions of any two test models and respective score formulas (or estimators) $x_i \equiv x_i(\mathbf{v})$, $i = 1, 2$. The initial subscripts refer to the respective test models and may be deleted when two possible score formulas or estimators $x_i(\mathbf{v})$, $i = 1, 2$, are considered in connection with

a single test model. In that case, the ratio

$$\text{RE}\ (\theta, x_1, x_2) = I(\theta, x_1)/I(\theta, x_2) \tag{20.7.1}$$

is called the *relative efficiency* (at θ) of x_1 to x_2. This function represents the relative precisions of estimators based on x_1 and x_2, respectively, being just the reciprocal of the ratio of the respective asymptotic variances at θ.

In the special cases where x_2 is such that $I(\theta, x_2) = I(\theta)$ for all θ, the ratio

$$\text{Eff}\ (\theta, x_1) = I(\theta, x_1)/I(\theta) \equiv I(\theta, x_1)/I(\theta, \hat{\theta}) \equiv \text{RE}\ (\theta, x_1, \hat{\theta}) \tag{20.7.2}$$

is called simply the *efficiency* (at θ) of x_1. We recall that $I(\theta, x_2) = I(\theta)$ when $x_2(\mathbf{v}) = \hat{\theta}(\mathbf{v})$, the maximum likelihood estimator, or when $x_2 = \sum_{g=1}^{n} a_g u_g$ in a logistic test model.

An estimator $\theta^* = \theta^*(\mathbf{v})$, for which Eff $(\theta, \theta^*) = 1$ for all θ, is called *efficient*. Examples are the maximum likelihood estimators under conditions indicated in Section 20.3 (or weaker conditions: see Cramér, 1946, pp. 498–506).

Returning to the general case of two possible test models for use in connection with a given latent trait θ, we now restrict consideration to one asymptotically efficient estimator based on each model, say $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. Then the ratio

$$\text{RP}\ (\theta) = I_1(\theta)/I_2(\theta) \equiv I_1(\theta, \hat{\theta}_1)/I_2(\theta, \hat{\theta}_2) \tag{20.7.3}$$

will be called the *relative precision* (at θ) *of the given respective test models.* This is just the ratio of asymptotic variances of maximum likelihood estimators $\sigma^2_{\hat{\theta}_2}/\sigma^2_{\hat{\theta}_1}$ at θ, since $\hat{\theta}_i$ is efficient and since therefore $\sigma^2_{\hat{\theta}_i} = 1/I(\theta)$.

Clearly RE (θ, x_1, x_2), the relative efficiency of two score formulas or estimators in one test model, and RP (θ), the relative precision of two test models, are special cases of the general *relative precision function*

$$\text{RP}\ (\theta) = I_1(\theta, x_1)/I_2(\theta, x_2), \tag{20.7.4}$$

which represents the relative precisions obtainable with any two estimators or scores x_1 and x_2, based, usually, on different test models.

Many examples are provided by the various test models, test designs, score formulas, and estimators discussed above. The following section is devoted to detailed consideration of the efficiency of the simple unweighted score formula in the general logistic test model.

20.8 Efficiency of Unweighted Scores in the Logistic Model

Using (20.2.7) and (20.4.7), we have the information function

$$I(\theta) = D^2 \sum_{g=1}^{n} a_g^2 \psi[DL_g(\theta)] \tag{20.8.1}$$

for any given logistic test model; and for $x(\mathbf{v}) = \sum_{g=1}^{n} a_g u_g$, we have $I(\theta, x) \equiv I(\theta, \hat{\theta}) \equiv I(\theta)$, for all θ. In the same model, for the unweighted score formula

$$\overline{u} = \frac{1}{n} \sum_{g=1}^{n} u_g,$$

we have

$$I(\theta, \overline{u}) \equiv \frac{[\partial \mathscr{E}(\overline{U} \mid \theta)/\partial\theta]^2}{\sigma^2(\overline{U}, \theta)} = \frac{\left\{D \sum_{g=1}^{n} a_g \psi[DL_g(\theta)]\right\}^2}{\sum_{g=1}^{n} \psi[DL_g(\theta)]}, \tag{20.8.2}$$

as we have shown in Section 17.7. Hence the efficiency of the unweighted score formula $\overline{u}$ is

$$\begin{aligned} \text{Eff } (\theta, \overline{u}) \equiv \frac{I(\theta, \overline{u})}{I(\theta)} &= \frac{\left\{\sum_{g=1}^{n} a_g \psi[DL_g(\theta)]\right\}^2}{\left\{\sum_{g=1}^{n} \psi[DL_g(\theta)]\right\} \left\{\sum_{g=1}^{n} a_g^2 \psi[DL_g(\theta)]\right\}} \\ &= \frac{\left(\sum_g a_g \psi_g\right)^2}{\left(\sum_g \psi_g\right)\left(\sum_g a_g^2 \psi_g\right)}. \end{aligned} \tag{20.8.3}$$

In the general case of unequal a_g, this function is less than unity for each θ.

To examine this function in some quantitative detail, we specialize our consideration to simple cases sharing some features with cases occurring in practice. Consider the special case of equal item difficulties, say $b_g \equiv b_1$, $g = 1, \ldots, n$, at ability level $\theta = b_1$. Since $\psi(0) = \frac{1}{4}$, we have

$$\text{Eff } (b_1, \overline{u}) = \frac{\left(\sum_{g=1}^{n} a_g\right)^2}{n\left(\sum_{g=1}^{n} a_g^2\right)} = \frac{\left((1/n)\sum_{g=1}^{n} a_g\right)^2}{(1/n)\sum_{g=1}^{n} a_g^2}. \tag{20.8.4}$$

Writing

$$\overline{a} = \frac{1}{n} \sum_{g=1}^{n} a_g, \qquad \overline{a^2} = \frac{1}{n} \sum_{g=1}^{n} a_g^2, \qquad \sigma_a^2 = \frac{1}{n} \sum_{g=1}^{n} (a_g - \overline{a})^2 \equiv \overline{a^2} - \overline{a}^2,$$

$$\text{CV } (a) \equiv \text{coefficient of variation among the } a_g = \sigma_a/\overline{a}, \tag{20.8.5}$$

we have

$$\text{Eff } (b_1, \overline{u}) = \frac{\overline{a}_g^2}{\overline{a^2}} = \frac{\overline{a}_g^2}{\overline{a}^2 + \sigma_a^2} = \frac{1}{1 + \sigma_a^2 \div \overline{a}^2} = \frac{1}{1 + \text{CV } (a)^2}. \tag{20.8.6}$$

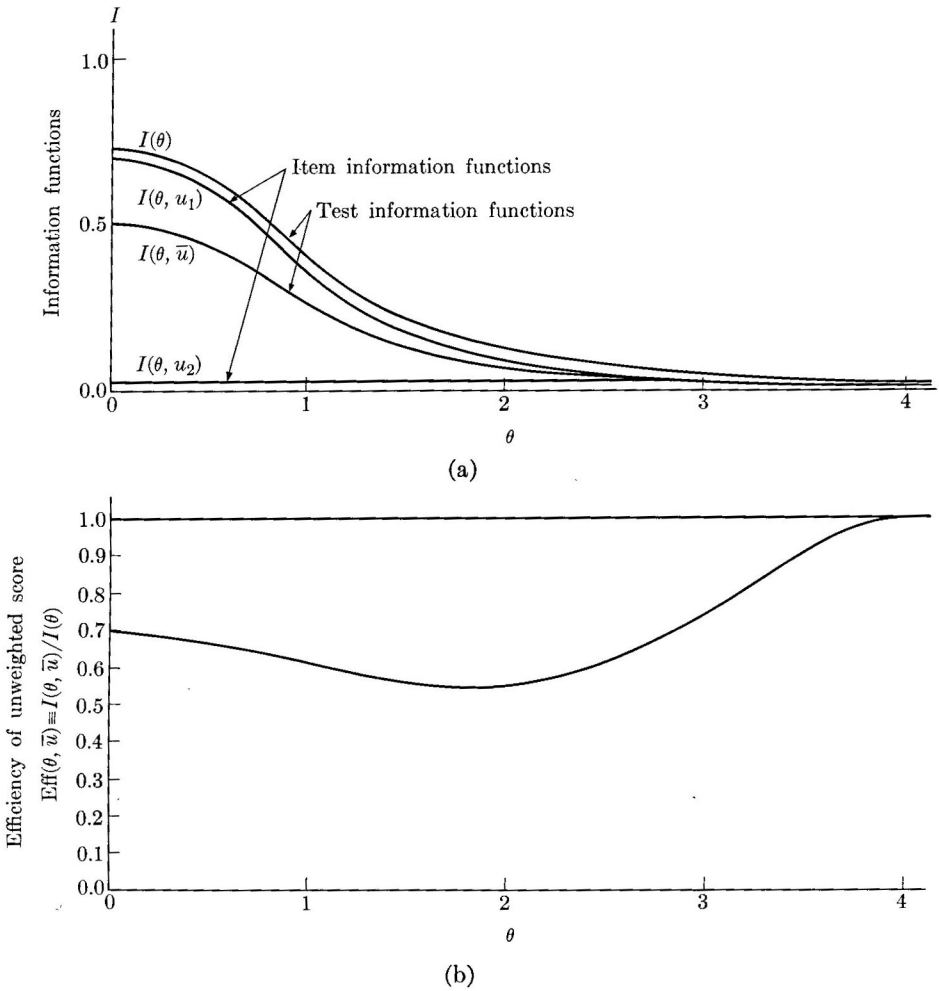

FIG. 20.8.1. Part (a) shows two item information functions, $I(\theta, \mu_1)$ and $I(\theta, \mu_2)$. It also shows two test information functions: $I(\theta)$, with optimal weighted score formula ($x = 0.98\,\mu_1 + 0.204\,\mu_2$), and $I(\theta, \mu)$, with unweighted formula ($x = \bar{\mu}$).

The function Eff $(\theta, \bar{u})$ is graphed in Figs. 20.8.1 through 20.8.4 for some cases in which $b_g = 0$ for all items. Figure 20.8.1a shows the information curves $I(\theta, u_1)$ and $I(\theta, u_2)$ of two *items*, with $a_1 = 0.7/\sqrt{0.51} \doteq 0.980$ and $a_2 = 0.2/\sqrt{0.96} \doteq 0.204$, respectively; these values represent extremes of the range of a_g encountered in practice. For comparison, the figure also includes the information curve $I(\theta) \equiv I(\theta, \hat{\theta}) \equiv I(\theta, x)$ of the hypothetical logistic test model containing just these two items, in which $x(\mathbf{v}) = a_1 u_1 + a_2 u_2 = 0.98 u_1 + 0.204 u_2$. By multiplying this function $I(\theta)$ by $n/2$, we obtain the information

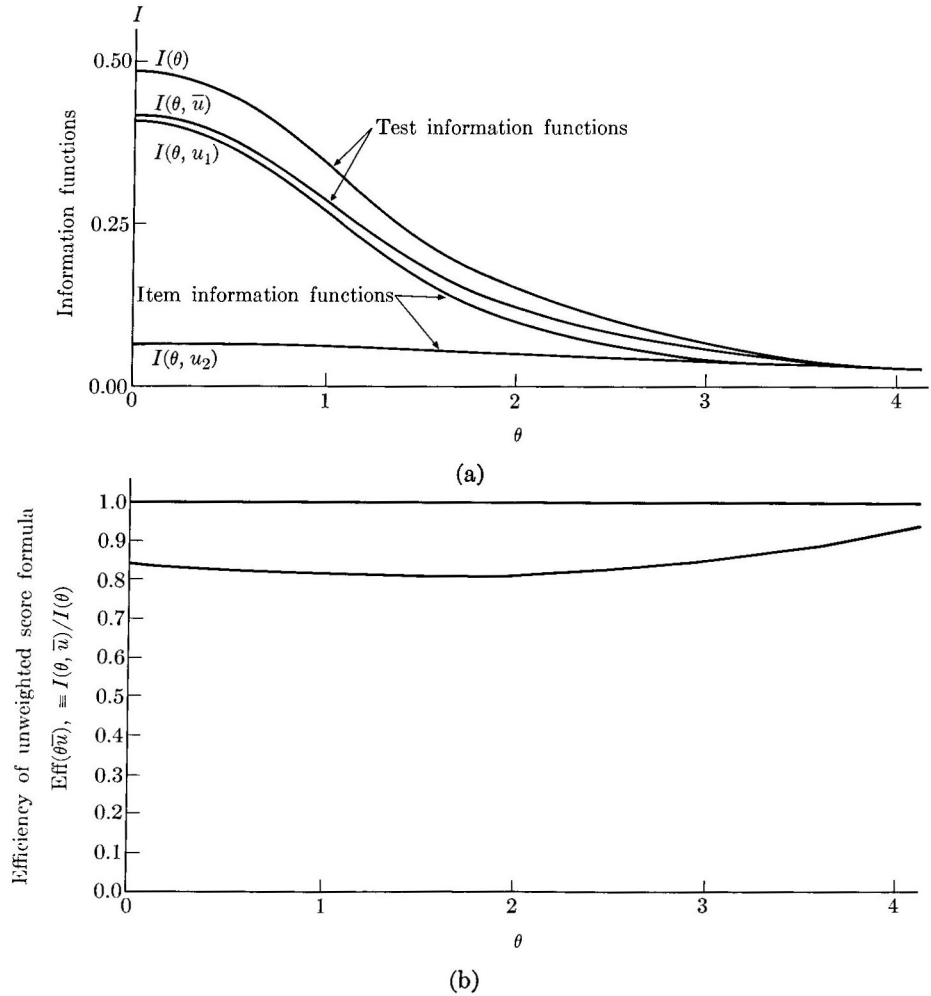

FIG. 20.8.2. Part (a) shows the two item information functions $I(\theta, \mu_1)$ and $I(\theta, \mu_2)$. It also shows the two test information functions $I(\theta)$ and $I(\theta, \mu)$, with unweighted and optimal weighted score formulas ($x = 0.75\mu_1 + 0.315\mu_2$).

function of a logistic test model consisting of n items, of which half have each of the extreme forms indicated. We see that the addition of item 2 to item 1 provides a small but appreciable gain, which is realized if the optimal weighted score formula is used.

For further comparison, the figure includes the information function $I(\theta, \bar{u})$ of the unweighted score based on the same hypothetical test of two items. When the latter is multiplied by $n/2$, we have the information function of $\bar{u}$ in a logistic test model with n items, half having each of the forms indicated.

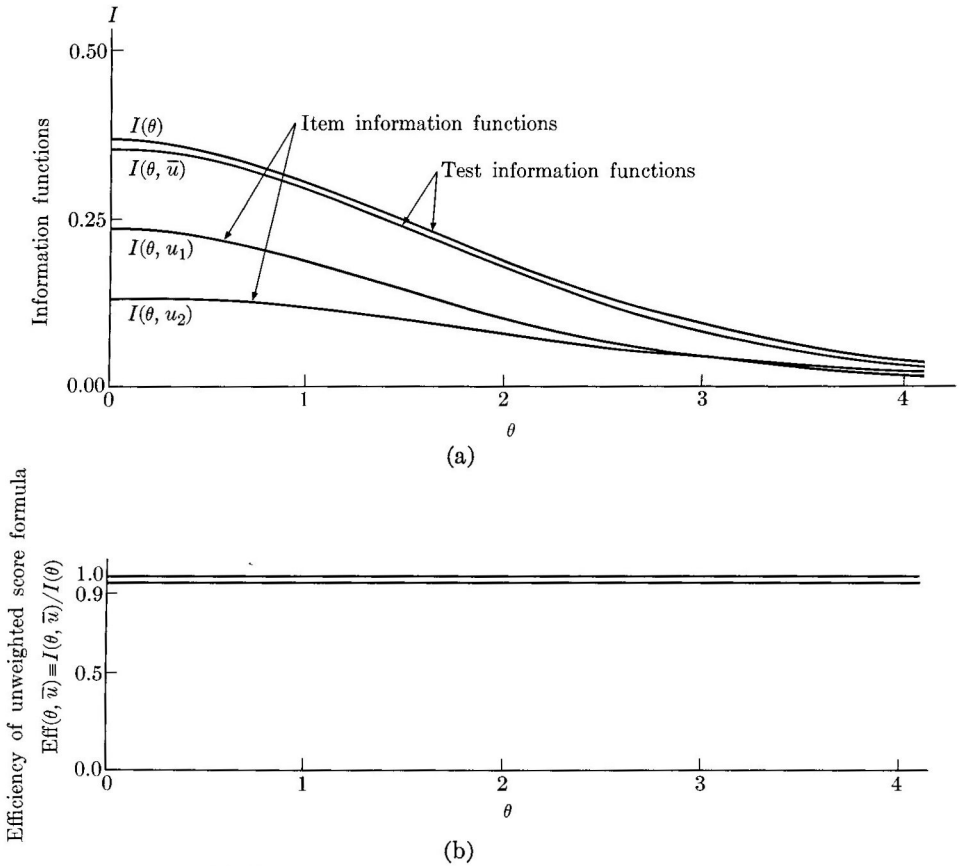

FIG. 20.8.3. Part (a) shows the two item information functions $I(\theta, \mu_1)$ and $I(\theta, \mu_2)$. It also shows the two test information functions $I(\theta)$ and $I(\theta, \mu)$, with unweighted and optimal weighted score formulas ($x = 0.578\,\mu_1 + 0.435\,\mu_2$).

We observe that adding the second (type of) item and using $\bar{u}$ causes an appreciable *reduction* in the values of the information function.

Figure 20.8.1b gives the efficiency

$$\text{Eff}\ (\theta, \bar{u}) \equiv I(\theta, \bar{u})/I(\theta) \tag{20.8.7}$$

of $\bar{u}$ in any logistic test model in which all the b_g are zero, and half the items have each of the a_g-values 0.98 and 0.204.

Figure 20.8.2 gives similar comparisons for items with $a_1 = 0.6/\sqrt{0.64} = 0.75$ and $a_2 = 0.3/\sqrt{0.91} \doteq 0.315$; and Fig. 20.8.3, similarly, for items with $a_1 = 0.5/\sqrt{0.75} \doteq 0.578$ and $a_2 = 0.4/\sqrt{0.84} \doteq 0.435$. We see that in the case represented by Fig. 20.8.3, where relative variation between the a_g is small,

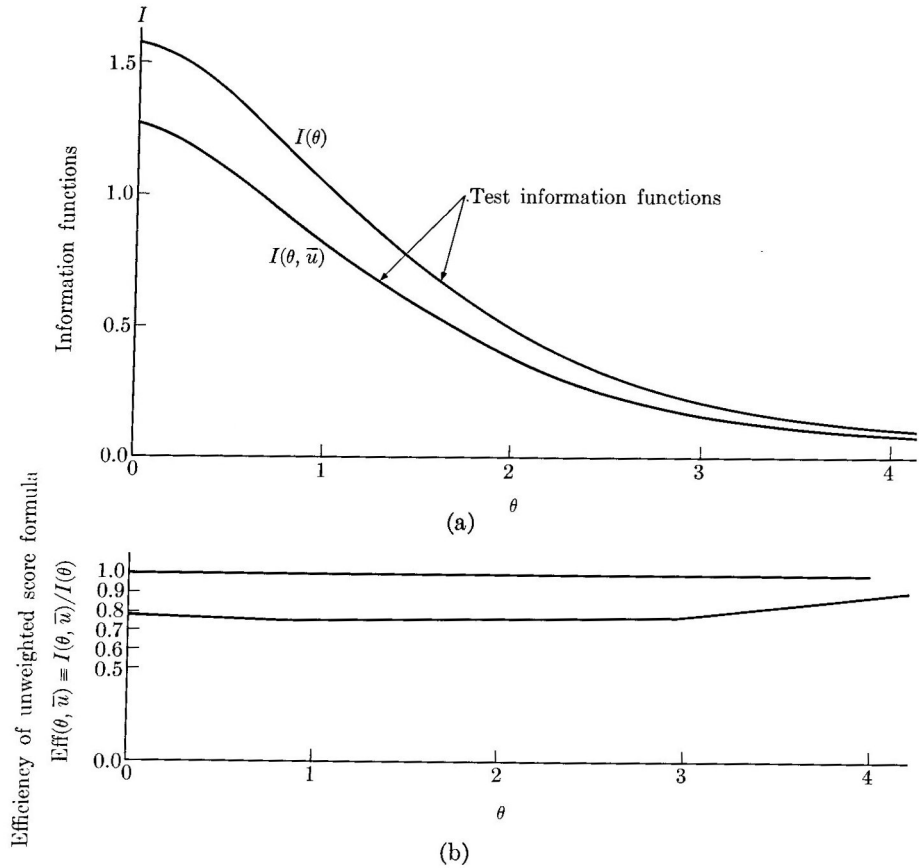

FIG. 20.8.4. Part (a) shows the two test item information functions $I(\theta)$ and $I(\theta, \mu)$, with unweighted and optimal weighted score formulas ($x = 0.98\,\mu_1 + 0.75\,\mu_2 + 0.578\,\mu_3 + 0.435\,\mu_4 + 0.315\,\mu_5 + 0.204\,\mu_6$).

the precision loss due to the use of unweighted scores is minor. In the intermediate case represented by Fig. 20.8.2, a major part of the information contributed by the second item is lost if the unweighted score is used. For the ability range ordinarily of interest, use of $\bar{u}$ rather than $x(\mathbf{v}) = \sum_{g=1}^{n} a_g u_g$ is equivalent, in a test such as that represented in Fig. 20.8.1, to discarding about one-third of the information available; in a test such as that represented in Fig. 20.8.2, this is equivalent to discarding about one-sixth of the information available; and, in Fig. 20.8.3, to about 3%.

Figure 20.8.4 gives the information function of a hypothetical logistic test model of six items, with all the b_g equal to zero, and with the a_g respectively equal to the six values represented above, namely, 1, 0.75, 0.578, 0.435, 0.315,

and 0.204; it also gives the information function of the unweighted score formula $\bar{u}$ in the same model, and the efficiency of $\bar{u}$, which is roughly 80% in the ability range usually of interest.

References and Selected Readings

BERKSON, J., Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, 1957, **13**, 28–34.

BIRNBAUM, A., Efficient design and use of tests of a mental ability for various decision-making problems. *Series Report No. 58–16.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, January 1957.

BIRNBAUM, A., On the estimation of mental ability. *Series Report No. 15.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (a)

BIRNBAUM, A., Further considerations of efficiency in tests of a mental ability. *Technical Report No. 17.* Project No. 7755–23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas, 1958. (b)

BIRNBAUM, A., Statistical theory of some quantal response models. *Annals of Mathematical Statistics*, 1958, **29**, 1284 (abstract). (c)

BIRNBAUM, A., Statistical theory of tests of a mental ability. *Op. cit.*, 1285 (abstract). (d)

BIRNBAUM, A., A unified theory of estimation, I. *Annals of Mathematical Statistics*, 1961, **32**, 112–135. (a)

BIRNBAUM, A., The theory of statistical inference. New York: Institute of Mathematical Sciences, New York University, 1961. (b) (Mimeographed)

BIRNBAUM, A., Statistical theory for logistic mental test models with a prior distribution of ability. *Research Bulletin 67–12.* Princeton, N.J.: Educational Testing Service, 1967.

CRAMÉR, H., *Mathematical methods of statistics.* Princeton, N.J.: Princeton University Press, 1946.

KENDALL, M. G., and A. STUART, *The advanced theory of statistics*, Vol. 2. London: Griffin, 1961.

LAWLEY, D. N., On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, **61-A**, 273–287.

LOEVINGER, JANE, The attenuation paradox in test theory. *Psychological Bulletin*, 1954, **51**, 493–504.

LORD, F. M., A theory of test scores. *Psychometric Monograph, No. 7.* Chicago: University of Chicago Press, 1950.

LORD, F. M., The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 1952, **17**, 181–194.

LORD, F. M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, 1953, **18**, 57–76.

SITGREAVES, ROSEDITH, A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), *Studies in item analysis and prediction.* Stanford: Stanford University Press, 1961.

SOLOMON, H., Probability and statistics in psychometric research: item analysis and classification techniques. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, **5,** pp. 169–184. Berkeley: University of California Press, 1956.

TUCKER, L. R, Maximum validity of a test with equivalent items. *Psychometrika,* 1946, **11,** 1–14.

WALD, A., Asymptotically shortest confidence intervals. *Annals of Mathematical Statistics,* 1942, **13,** 127–137.