Research Report

The Information a Test Provides on an Ability Parameter

Shelby J. Haberman

Research & Development

May 2007
RR-07-18

The Information a Test Provides on an Ability Parameter

Shelby J. Haberman
ETS, Princeton, NJ

May 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

In item-response theory, if a latent-structure model has an ability variable, then elementary information theory may be employed to provide a criterion for evaluation of the information the test provides concerning ability. This criterion may be considered even in cases in which the latent-structure model is not valid, although interpretation of the criterion is more complex in this case. It is also possible to consider reliability results for inferences about the ability variable. Use of sampling to develop required parameter estimates is straightforward. Applications of the criterion in zero-, one-, and two-parameter logistic (0PL, 1PL, and 2PL) examples are provided.

Key words: Latent variable, entropy, reliability

Acknowledgments

The author would like to thank Paul Holland for very helpful discussions.

Given an item-response model, information theory can be employed to assess the information concerning an ability parameter that is provided by the data. The argument is the same as the argument used to define the information provided by an experiment (Lindley, 1956), although some care must be taken to treat the issue of false models. In this paper, a random ability parameter is generated that corresponds to a conventional ability parameter if a model is true but is shown to be definable and estimable even if the model is not valid. In Section 1, the basic results from information theory are provided, and suitable estimation procedures are developed. In addition, the zero-, one-, and two-parameter logistic (0PL, 1PL, and 2PL) models to be examined are described for the case of a model with a normal ability distribution and for a model with a multinomial ability distribution. Section 2 summarizes results obtained by analysis of a 45-item right-scored multiple-choice writing examination from the *Praxis*^{TM} series that has previously been employed (Haberman, 2005) to illustrate use of latent-structure models. This example is quite suitable for analysis in that the data are easily shown not to satisfy any standard item response model, but standard 1PL models appear to provide reasonable approximations of the observed distribution of item responses for the 8,686 examinees. Section 3 provides some conclusions. The results on the whole suggest that efforts to estimate ability parameters are surprisingly successful even when models are not valid.

1 Information Concerning Ability

In this section, information theory is used to examine the information provided concerning ability variables associated with item-response models. Consider binary random item responses X_{vi} by examinee v to item i, $1 \le i \le I$, $1 \le v \le n$, where the number I of items is at least 2 and the number n of examinees is at least 2. For each examinee v, $1 \le v \le n$, and each item i, $1 \le i \le I$, let X_{vi} have possible integer values 0 or 1, where X_{vi} is 1 if a correct response is given to by examinee v to item i and X_{vi} is 0 otherwise. The responses for examinee v may be described by $\mathbf{X}_v$, the I-dimensional response vector with coordinates X_{vi}, $1 \le i \le I$. If $\mathcal{X}$ denotes the set of I-dimensional vectors $\mathbf{x}$ with all coordinates 0 or 1, then $\mathcal{X}$ is the set of possible values of $\mathbf{X}_v$. It is assumed in this paper that the probability $p(\mathbf{x})$ that $\mathbf{X}_v = \mathbf{x}$ is positive for each possible $\mathbf{x}$ in $\mathcal{X}$, and it is assumed that the response vectors $\mathbf{X}_v$, $1 \le v \le n$, are mutually independent and identically distributed. To simplify notation, the convention is adopted that $\mathbf{X}_1$ is denoted by $\mathbf{X}$, and X_{1i} is denoted by X_i.

1.1 Latent-Structure Models

This paper is concerned with the use and interpretation of latent ability vectors for item responses when the models are not necessarily true. In the models under study, a single latent variable is used; however, generalization to multiple latent variables is straightforward. Illustrations are based on the 2PL model (Hambleton, Swaminathan, & Rogers, 1991), but the methods proposed apply much more widely. The methodology proposed can be employed both with ability variables with a normal distribution or with ability variables with a multinomial distribution.

For any model under study, it is assumed that some real latent variable θ_v is associated with each examinee v. It is assumed in each model under consideration that the pairs $(\mathbf{X}_v, \theta_v)$, $1 \leq v \leq n$, are mutually independent and identically distributed, and the local independence assumption is made that the responses X_{vi}, $1 \leq i \leq I$, of subject v are conditionally independent given θ_v. The convention is adopted that θ_1 is denoted by θ. Associated with each item i is an item characteristic curve (ICC) P_i such that $P_i(\omega)$ is the probability that $X_i = 1$ given that $\theta = \omega$. In all models in this paper, the 2PL assumption is made that

$$P_i(\omega) = [1 + \exp(-a_i\omega + \gamma_i)]^{-1} \tag{1}$$

for an unknown positive real a_i, the item discrimination of item i, and for a real unknown γ_i. The ratio $\beta_i = \gamma_i/a_i$ is the item difficulty for item i. In the 1PL (Rasch) models under study, each a_i is assumed equal. In the 0PL (binomial) models under study, each a_i is assumed equal and each γ_i is assumed equal. In the normal latent-variable models under study, θ is assumed to have a standard normal distribution. In the latent-class models under study, an integer $K \geq 2$ and distinct real numbers τ_k, $1 \leq k \leq K$, are specified, θ_v can only assume the values τ_k, $1 \leq k \leq K$, and $\theta_v = \tau_k$ with an unknown but positive probability.

In any of these latent-variable models, if the model is correct and the model parameters are known, then Bayes' theorem may be employed for inferences about the latent variable θ_v based on the response vector $\mathbf{X}_v$ (Bock & Aitkin, 1981). In practice, maximum-likelihood estimates of model parameters may be used instead of the model parameters themselves if the sample size is large.

If the model is not correct, then the meaning of inferences concerning the θ_v is not obvious. Indeed, it is not even obvious what the θ_v are. In this paper, a procedure based on information

theory is employed to construct random variables θ_v with conditional distributions given the responses $\mathbf{X}_v$ that are of the same form as the corresponding conditional distributions obtained if the model holds. These random variables θ_v may be constructed by use of a random number generator capable of providing independently distributed uniform random numbers U_v for $1 \leq v \leq n$, where the U_v are independent of the observed responses X_{iv} and have range $(0, 1)$. Information theory is used to evaluate the strength of the relationship of these θ_v to the observed examinee responses $\mathbf{X}_v$ of examinee v and to evaluate the reliability of a standard estimate of θ_v. Maximum likelihood is readily used to estimate all needed parameters and measures.

1.2 Models and Information Theory

To develop the variables θ_v, information theory is employed to obtain approximations to the response probabilities $p(\mathbf{x})$, $\mathbf{x}$ in $\mathcal{X}$ that are consistent with the latent-variable model under study (Gilula & Haberman, 1994, 1995; Savage, 1971; Shannon, 1948). To apply information theory, some preliminary definitions are needed. Let $\mathbf{p}$ denote the function on $\mathcal{X}$ with value $p(\mathbf{x})$ for $\mathbf{x}$ in $\mathcal{X}$, so that $\mathbf{p}$ determines the distribution of each response vector $\mathbf{X}_v$. Let S be the simplex of nonnegative real functions $\mathbf{r}$ on $\mathcal{X}$ with sum $\sum_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x}) = 1$, so that $\mathbf{p}$ is in S and to any $\mathbf{r}$ in S corresponds a random vector $\mathbf{Y}$ with values in $\mathcal{X}$ such that $\mathbf{Y} = \mathbf{x}$ with probability $r(\mathbf{x})$ for each $\mathbf{x}$ in $\mathcal{X}$. For the model under study, let T be the set such that $\mathbf{p}$ is in T if, and only if, the model holds. For example, in the case of a normal 2PL model, $\mathbf{r}$ is in the set T if, and only if, for each $\mathbf{x}$ in $\mathcal{X}$,

$$r(\mathbf{x}) = \int r(\mathbf{x}|\omega)\phi(\omega)d\omega, \tag{2}$$

where

$$r(\mathbf{x}|\omega) = \prod_{i=1}^{I} [P_i(\omega)]^{x_i}[1 - P_i(\omega)]^{1-x_i}, \tag{3}$$

ϕ is the density function of the standard normal distribution, and (1) holds for some real $a_i > 0$ and real γ_i, $1 \leq i \leq I$. Observe that if the normal 2PL model holds for pairs $(\mathbf{X}_v, \theta_v)$ and $\mathbf{r} = \mathbf{p}$, then it is clear that θ_v has a standard normal distribution and (1), (2), and (3) hold for some real $a_i > 0$ and γ_i, so that $\mathbf{p}$ is in T.

On the other hand, if $\mathbf{p}$ is in T, then (1), (2), and (3) hold for $\mathbf{r} = \mathbf{p}$ for some real $a_i > 0$ and real γ_i. Given the uniform random numbers U_v, $1 \leq v \leq n$, random variables θ_v may be constructed so that the pairs $(\mathbf{X}_v, \theta_v)$ are mutually independent and the conditional density of θ_v

given $\mathbf{X}_v = \mathbf{x}$ is

$$f(\omega|\mathbf{x}) = [p(\mathbf{x})]^{-1}\phi(\omega)r(\mathbf{x}|\omega) \tag{4}$$

for ω in R. It suffices to define the strictly increasing and continuous distribution function

$$F(\omega|\mathbf{x}) = \int_{-\infty}^{\omega} f(\nu|\mathbf{x})d\nu \tag{5}$$

for ω real and to let

$$F(\theta_v|\mathbf{X}_v) = U_v \tag{6}$$

(Rao, 1973, p. 87). The marginal density of θ_v is

$$\sum_{\mathbf{x}\in\mathcal{X}} f(\omega|\mathbf{x})p(\mathbf{x}) = \phi(\omega),$$

so that θ_v has a standard normal distribution. Bayes' theorem implies that the conditional probability that $\mathbf{X}_v = \mathbf{x}$ given $\theta_v = \omega$ is $r(\mathbf{x}|\omega)$. Thus the resulting pairs $(\mathbf{X}_v, \theta_v)$ satisfy the normal 2PL model.

Similar arguments apply to other models under study. For example, in the 2PL latent-class case, $\mathbf{r}$ is in T if, and only if, (1) holds for some $a_i > 0$ and real γ_i,

$$r(\mathbf{x}) = \sum_{k=1}^{K} \pi_k r(\mathbf{x}|\tau_k)$$

for $\pi_k > 0$, $1 \le k \le K$, such that $\sum_{k=1}^{K} \pi_k = 1$, and (3) holds for ω real. If the latent-class 2PL model holds for pairs $(\mathbf{X}_v, \theta_v)$ and $\mathbf{r} = \mathbf{p}$, then $\theta_v = \tau_k$ with probability $\pi_k > 0$, $1 \le k \le K$, where $\sum_{k=1}^{K} \pi_k = 1$, and (1), (2), and (3) hold for some real $a_i > 0$ and γ_i, so that $\mathbf{p}$ is in T.

On the other hand, let $\mathbf{p}$ be in the set T defined for the latent-class 2PL model. Then (1), (2), and (3) hold for $\mathbf{r} = \mathbf{p}$ for some real $a_i > 0$ and real γ_i and for some $\pi_k > 0$, $1 \le k \le K$, such that $\sum_{k=1}^{K} \pi_k = 1$. Construct a random variable θ_v from the uniform random number U_v such that the conditional probability that $\theta_v = \tau_k$ given $\mathbf{X}_v = \mathbf{x}$ is equal to

$$f(\tau_k|\mathbf{x}) = [p(\mathbf{x})]^{-1}\pi_k r(\mathbf{x}|\tau_k) \tag{7}$$

for $1 \le k \le K$. If the τ_k are increasing in k, then this construction is accomplished by letting $\theta_v = k$ by use of the distribution function values

$$F(\tau_k|\mathbf{x}) = \sum_{h=1}^{k} f(\tau_h|\mathbf{x}) \tag{8}$$

for $1 \leq k \leq K$. Let $\theta_v = \tau_k$ for $1 \leq k \leq K$ if $U_v \leq F(\tau_k|\mathbf{X}_v)$ and either $k = 1$ or $U_v > F(\tau_{k-1}|\mathbf{X}_v)$. The probability that $\theta_v = \tau_k$ is π_k for $1 \leq k \leq K$, and the conditional probability that $\mathbf{X}_v = \mathbf{x}$ given $\theta_v = \tau_k$ is $r(\mathbf{x}|\tau_k)$.

To examine the more general case in which the model under study need not be true, consider probability prediction of $\mathbf{X}_v$ by use of a logarithmic penalty function (Gilula & Haberman, 1994, 1995, 2000). If the prediction $\mathbf{r}$ in S is employed and if the penalty is $-\log r(\mathbf{x})$ for $\mathbf{X}_v = \mathbf{x}$ in $\mathcal{X}$, then the expected penalty per item is

$$J(\mathbf{r}) = -I^{-1}E(\log r(\mathbf{X})) = -I^{-1}\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \log r(\mathbf{x}),$$

where $0 \log 0 = 0$ (Haberman, 2005). The minimum of $J(\mathbf{r})$) is achieved if, and only if, $\mathbf{r} = \mathbf{p}$, and

$$H(\mathbf{X}) = J(\mathbf{p}) = -I^{-1}\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x})$$

is the entropy per item of $\mathbf{X}$. To measure the discrepancy between $\mathbf{p}$ and the probability model that $\mathbf{p}$ is in T, consider the minimum $H^*(\mathbf{X})$ of $J(\mathbf{r})$ for $\mathbf{r}$ in T. This minimum cannot be less than $H(\mathbf{X})$, and $H^*(\mathbf{X}) = H(\mathbf{X})$ if the model holds. Thus a measure of model error is the minimum discriminant information $D(\mathbf{X}) = H^*(\mathbf{X}) - H(\mathbf{X})$ for distinguishing between the true probability vector $\mathbf{p}$ and a probability vector $\mathbf{r}$ in T (Haberman, 1982; Kullback, 1968). The measure $D(\mathbf{X})$ is nonnegative. If $D(\mathbf{X})$ is positive, then the model is not valid. One may regard $D(\mathbf{X})$ as a measure of the average increase per item in log penalty due to use of the model that $\mathbf{p}$ is in T rather than some model that is correct.

If $\mathbf{p}^*$ in T satisfies $J(\mathbf{p}^*) = H^*(\mathbf{X})$, then $\mathbf{p}^*$ is an optimal approximation to $\mathbf{p}$ within T according to the criterion of expected log penalty. If the model holds, then $\mathbf{p}^* = \mathbf{p}$. In general, $\mathbf{p}^*$ satisfies conditions quite similar to maximum-likelihood equations for log-linear models in the case of indirect observation (Haberman, 1979, chap. 10) or maximum-likelihood equations for exponential families in which observation is incomplete (Dempster, Laird, & Rubin, 1977). These conditions may be described in terms of moments that involve a vector $\mathbf{X}^*$ with the same possible values as the observed vector $\mathbf{X}$ and a random variable θ^*. As required in a latent-structure model, the coordinates X_i^* of $\mathbf{X}^*$ are conditionally independent given θ^*, and $\mathbf{X}^* = \mathbf{x}$ with probability $\mathbf{p}^*(\mathbf{x})$. Additional moment restrictions and distribution restrictions are imposed on θ^* and $\mathbf{X}^*$ by the specific model under consideration. In the case of the normal 2PL model, for each item i, positive a_i and real γ_i exist such that the conditional probability that $X_i^* = 1$ given $\theta^* = \omega$

is $P_i(\omega)$, where $P_i(\omega)$ satisfies (1). In addition, θ^* has a standard normal distribution. For each item i, the moment conditions are added that X_i^* and X_i have the same expected values, and the expected value $E(X_i^*\theta^*)$ is equal to the expected value of $E(X_i\theta)$, where θ is a random variable such that the conditional distribution of θ given $\mathbf{X} = \mathbf{x}$ is the same as the conditional distribution of θ^* given $\mathbf{X}^* = \mathbf{x}$ for each $\mathbf{x}$ in $\mathcal{X}$.

In the latent-class 2PL model, it remains the case that, for each item i, positive a_i and real γ_i exist such that the conditional probability that $X_i^* = 1$ given $\theta^* = \omega$ is $P_i(\omega)$, where $P_i(\omega)$ satisfies (1); however, θ^* now assumes only the values τ_k, $1 \leq k \leq K$, and, for $1 \leq k \leq K$, the probability is positive that $\theta^* = \tau_k$. It also remains the case that, for each item i, X_i^* and X_i have the same expected value, and $E(X_i^*\theta^*) = E(X_i\theta)$, where the conditional distribution of θ given $\mathbf{X} = \mathbf{x}$ is the same as the conditional distribution of θ^* given $\mathbf{X}^* = \mathbf{x}$ for each $\mathbf{x}$ in $\mathcal{X}$. The added condition is imposed that θ^* and θ have the same distribution.

In all cases, the ability variables θ_v under study will be functions of the uniform random variable U_v and the observed responses $\mathbf{X}_v$, and the conditional distribution of θ_v given $\mathbf{X}_v = \mathbf{x}$ will be the same as the conditional distribution of θ^* given $\mathbf{X}^* = \mathbf{x}$. If the proposed model actually holds, then $(\theta_v, \mathbf{X}_v)$ will have the same joint distribution as $(\theta, \mathbf{X})$. In the normal 2PL example, the θ_v are still obtained by use of (1), (2), (3), (4), (5), and (6). In the latent-class 2PL example, the θ_v are still obtained by use of (1), (2), (3), (7), and (8).

1.3 Evaluation of the Ability Variable

The variable θ_v constructed for each examinee v need not satisfy the local independence requirement of the latent-structure model if the latent-structure model itself does not hold. Nonetheless, θ_v may still be an effective predictor of the response vector $\mathbf{X}_v$. The effectiveness of θ_v may be evaluated by use of information theory in conjunction with the latent-structure model. Given θ_v, the appropriate probability prediction of $\mathbf{X}_v$ is the conditional probability function $\mathbf{p}^*(\theta_v)$ on $\mathcal{X}$ with value $p^*(\mathbf{x}|\theta_v)$ at $\mathbf{x}$ in $\mathcal{X}$. The expected log penalty per item is

$$H^*(\mathbf{X}|\theta) = -I^{-1}E(\log p^*(\mathbf{X}|\theta)).$$

If the model holds, then $H^*(\mathbf{X}|\theta)$ is the conditional entropy per item $H(\mathbf{X}|\theta)$ of $\mathbf{X}$ given θ. Given conditional independence of the X_i^* given θ^*,

$$H^*(\mathbf{X}|\theta) = I^{-1}\sum_{i=1}^{I} H^*(X_i|\theta),$$

where

$$H^*(X_i|\theta) = -E(X_i \log P_i(\theta) + (1 - X_i) \log[1 - P_i(\theta)]).$$

If the model holds, then $H^*(X_i|\theta)$ is the conditional entropy $H(X_i|\theta)$ of response X_i given the variable θ.

As shown in this section,

$$H^*(\mathbf{X}|\theta) < H^*(\mathbf{X}). \tag{9}$$

The difference

$$\Delta(\mathbf{X}|\theta) = H^*(\mathbf{X}) - H^*(\mathbf{X}|\theta) \tag{10}$$

provides a measure of the information per item on $\mathbf{X}$ provided by θ.

1.4 The Latent-Class Case

To verify (9), it is helpful to distinguish between the latent-class models and the normal models. In the latent-class models, the logarithmic penalty arguments for $\mathbf{X}$ can also be applied to $(\mathbf{X}, \theta)$ and θ. Let $f^(\tau_k|\mathbf{x})$ denote the conditional probability that $\theta^* = \tau_k$ given that $\mathbf{X}^* = \mathbf{x}$, so that

$$p^*(\mathbf{x}, \tau_k) = p^*(\mathbf{x}) f^*(\tau_k|\mathbf{x})$$

is the joint probability that $(\mathbf{X}^*, \theta^*) = (\mathbf{x}, \tau_k)$ for $\mathbf{x}$ in $\mathcal{X}$ and $1 \leq k \leq K$. The expected log penalty per item for the corresponding probability prediction for $(\mathbf{X}, \theta)$ is then

$$H^*(\mathbf{X}, \theta) = -I^{-1} E(\log p^*(\mathbf{X}, \theta)). \tag{11}$$

If the model holds, then $H^*(\mathbf{X}, \theta)$ is just the joint entropy $H(\mathbf{X}, \theta)$ of $\mathbf{X}$ and θ, so that $H^*(\mathbf{X}, \theta) \geq H(\mathbf{X}, \theta)$, with equality if the model holds.

The probability prediction for $\theta = \tau_k$ given that $\mathbf{X} = \mathbf{x}$ is $f^*(\tau_k|\mathbf{x})$, so that the expected log penalty for prediction of θ by $\mathbf{X}$ is

$$H^*(\theta|\mathbf{X}) = -E(\log f^*(\theta|\mathbf{X})). \tag{12}$$

In this instance, the definition of θ implies that $H^*(\theta|\mathbf{X})$ is also the conditional entropy $H(\theta|\mathbf{X})$ of θ given $\mathbf{X}$. It follows that

$$H^*(\mathbf{X}, \theta) = H^*(\mathbf{X}) + I^{-1} H(\theta|\mathbf{X}). \tag{13}$$

An alternative decomposition to (13) may be created by use of the marginal distribution of θ predicted by the model. Let $\theta^* = \tau_k$ with probability π_k^*. Then

$$p^*(\mathbf{x}, \tau_k) = \pi_k^* p^*(\mathbf{x}|\tau_k).$$

Consider the probability prediction of $\theta = \tau_k$ by π_k^*. Observe that the actual probability that $\theta = \tau_k$ is $\pi_k = E(f^*(\tau_k|\mathbf{X}))$. The expected log penalty is then

$$H^*(\theta) = -E(\log \pi_\theta^*) = -\sum_{k=1}^{K} \pi_k \log \pi_k^*, \tag{14}$$

and the entropy of θ is

$$H(\theta) = -\sum_{k=1}^{K} \pi_k \log \pi_k.$$

Thus $H^*(\theta) \geq H(\theta)$, with equality if the model holds.

A new decomposition of the joint expected log penalty $H^*(\mathbf{X}, \theta)$ is now available, for

$$H^*(\mathbf{X}, \theta) = H^*(\mathbf{X}|\theta) + I^{-1} H^*(\theta). \tag{15}$$

By (13) and (15),

$$\Delta(\mathbf{X}|\theta) = I^{-1} \Delta(\theta|\mathbf{X}), \tag{16}$$

where

$$\Delta(\theta|\mathbf{X}) = H^*(\theta) - H(\theta|\mathbf{X}). \tag{17}$$

Because $H^*(\theta) \geq H(\theta)$ and $H(\theta) \geq H(\theta|\mathbf{X})$ (Shannon, 1948), it follows that $\Delta(\mathbf{X}|\theta)$ and $\Delta(\theta|\mathbf{X})$ are nonnegative. The constraint that $a_i > 0$ and the constraint that each $\pi_k > 0$ imply that θ and $\mathbf{X}$ are not independent, so that $H(\theta) > H(\theta|\mathbf{X})$. Thus $\Delta(\mathbf{X}|\theta)$ and $\Delta(\theta|\mathbf{X})$ are positive.

The measures $\Delta(\mathbf{X}|\theta)$ and $\Delta(\theta|\mathbf{X})$ indicate the strength of the relationship between θ and $\mathbf{X}$. The perspectives are a bit different. In the case of $\Delta(\mathbf{X}|\theta)$, one considers the improvement per item in probability prediction of $\mathbf{X}$ achieved by use of θ. In the case of $\Delta(\theta|\mathbf{X})$, one considers the improvement in probability prediction of θ by use of the response vector $\mathbf{X}$.

1.5 The Normal Case

Verification of (9) in the normal case is a bit different due to the continuity of θ. Some penalties based on log probabilities are replaced by log densities (Gilula & Haberman, 2000), and

penalties are not necessarily nonnegative. The conditional density $f^*(\omega|\mathbf{x})$ of θ^* given $\mathbf{X}^* = \mathbf{x}$ is considered. The product

$$p^*(\mathbf{x}, \omega) = p^*(\mathbf{x}) f^*(\omega|\mathbf{x})$$

is used for prediction of $(\mathbf{X}, \theta)$ for $\mathbf{x}$ in $\mathcal{X}$ and ω real, and the log penalty per item for the prediction is

$$H^*(\mathbf{X}, \theta) = -I^{-1} E(\log p^*(\mathbf{X}, \theta)). \tag{18}$$

The density prediction for θ given $\mathbf{X} = \mathbf{x}$ is the function with value $f^*(\theta|\mathbf{x})$ at θ real, and the expected log penalty for prediction of θ by $\mathbf{X}$ is still given by (12) and (13) still holds. The expected log penalty of (12) is also the information of θ given $\mathbf{X}$.

Once again an alternative decomposition to (13) employs the marginal distribution of θ predicted by the model. For the conditional probability $p^*(\mathbf{x}|\omega)$ that $\mathbf{X}^* = \mathbf{x}$ given that the standard normal random variable $\theta^* = \omega$, one has the decomposition

$$p^*(\mathbf{x}, \omega) = \phi(\omega) p^*(\mathbf{x}|\omega).$$

Consider density prediction of θ by ϕ. The actual density f of θ satisfies

$$f(\omega) = E(f^*(\omega|\mathbf{X})) \tag{19}$$

for ω real, so that the expected log penalty is

$$H^*(\theta) = -E(\log \phi(\theta)) = -\int f(\omega) \log \phi(\omega) d\omega = [\log(2\pi) + E(\theta^2)]/2, \tag{20}$$

and the information of θ is

$$H(\theta) = -\int f(\omega) \log f(\omega) d\omega,$$

so that $H^*(\theta) \geq H(\theta)$, with equality if the model holds. Once again, (15) holds, so that (16) and (17) follow. Essentially the same argument suffices to show that $\Delta(\mathbf{X}|\theta)$ and $\Delta(\theta|\mathbf{X})$ are positive.

For the normal case, there is some necessary relationship between the measure $\Delta(\theta|\mathbf{X})$ and the conditional variance $\sigma(\theta|\mathbf{x})$ of θ given $\mathbf{X} = \mathbf{x}$. A standard information inequality implies that

$$2\Delta(\theta|\mathbf{X}) \geq E(\theta^2) - 1 - E(\log \sigma^2(\theta^2|\mathbf{X}))$$

(Rao, 1973, pp. 162–163). If the model holds, then $E(\theta^2)$ is 1.

1.6 Mean Squared Error

Given either the latent class or normal case, the latent variable θ can also be evaluated in terms of mean squared error of prediction. One can compute the conditional expected value $E(\theta|\mathbf{X})$ of θ given $\mathbf{X}$ to obtain the expected a posteriori (EAP) estimate of θ (Bock & Aitkin, 1981). The proportional reduction of mean squared error from prediction of θ by $E(\theta|\mathbf{X})$ rather than by $E(\theta)$ is then

$$\rho^2(\theta|\mathbf{X}) = 1 - E(\sigma^2(\theta|\mathbf{X}))/\sigma^2(\theta).$$

Computations can be simplified in practice by considering the weighted sum

$$W(\mathbf{x}) = \sum_{i=1}^{I} a_i x_i$$

for $\mathbf{x}$ in $\mathcal{X}$. In all models, $E(\theta|\mathbf{x})$ is a function of $W(\mathbf{x})$. Let

$$V(\omega) = \prod_{i=1}^{I} [1 + \exp(a_i\omega - \gamma_i)]^{-1}.$$

In the normal case,

$$E(\theta|\mathbf{x}) = \frac{\int \theta \exp[W(\mathbf{x})\omega] V(\omega)\phi(\omega)d\omega}{\int \exp(W(\mathbf{x})\omega) V(\omega)\phi(\omega)d\omega}.$$

Numerical computations can be performed efficiently via adaptive Gauss-Hermite quadrature (Haberman, 2006) with 9 integration points. In the latent-class case,

$$E(\theta|\mathbf{x}) = \frac{\sum_{k=1}^{K} \tau_k \exp[W(\mathbf{x})\tau_k] V(\tau_k)\pi_k^*}{\sum_{k=1}^{K} \exp[W(\mathbf{x})\tau_k] V(\tau_k)\pi_k^*}.$$

Similar arguments apply to conditional variances.

1.7 Estimation

In practice, maximum-likelihood estimation may be used to treat the problem that the distribution of $\mathbf{X}$ is not known. Given use of maximum likelihood, estimates $\hat{a}_i$ of a_i and $\hat{\gamma}_i$ of γ_i are obtained for the appropriate model. In the latent-class case, maximum-likelihood estimates $\hat{\pi}_k$ of π_k are used to estimate π_k^*. In all cases, $W(\mathbf{x})$ is approximated by

$$\hat{W}(\mathbf{x}) = \sum_{i=1}^{I} \hat{a}_i x_i,$$

and $V(\omega)$ is approximated by

$$\hat{V}(\omega) = \prod_{i=1}^{I} [1 + \exp(\hat{a}_i\omega - \hat{\gamma}_i)]^{-1}.$$

The log likelihood is used for estimation of $H^*(\mathbf{X})$. Here the log likelihood function is

$$\ell(\mathbf{r}) = \sum_{v=1}^{n} \log r(\mathbf{X}_v)$$

for $\mathbf{r}$ in S. For the subset T of S corresponding to the model under study, the standard estimate of $H^*(\mathbf{X})$ is

$$\hat{H}(\mathbf{X}) = (nk)^{-1}\ell(T),$$

where $\ell(T)$ is the supremum of $\ell(\mathbf{r})$ for $\mathbf{r}$ in T (Gilula & Haberman, 1994, 1995).

In other calculations, $\mathbf{p}^*$ is normally approximated by a maximum-likelihood estimate $\hat{\mathbf{p}}$ in T such that $\ell(\hat{\mathbf{p}}) = \ell(T)$, and $\mathbf{p}$ is approximated by $\bar{\mathbf{p}}$ for $\bar{p}(\mathbf{x})$ equal to the fraction of v, $1 \leq v \leq n$, such that $\mathbf{X}_v = \mathbf{x}$. One approximates $P_i(\omega)$ by

$$\hat{P}_i(\omega) = [1 + \exp(-\hat{a}_i\omega + \hat{\gamma}_i)]^{-1},$$

so that $p^*(\mathbf{x}|\omega)$ is approximated by

$$\hat{p}(\mathbf{x}|\omega) = \prod_{i=1}^{I} [\hat{P}_i(\omega)]^{x_i}[1 - \hat{P}_i(\omega)]^{1-x_i}.$$

In the normal case, $p(\mathbf{x})$ is then approximated by

$$\hat{p}(\mathbf{x}) = \int \hat{p}(\mathbf{x}|\omega)\phi(\omega)d\omega,$$

$f^*(\omega|\mathbf{x})$ is approximated by

$$\hat{f}(\omega|\mathbf{x}) = \hat{p}(\mathbf{x}|\omega)\phi(\omega)/\hat{p}(\mathbf{x}),$$

$E(\theta|\mathbf{x})$ is approximated by

$$\hat{E}(\theta|\mathbf{x}) = \frac{\int \omega \exp[\hat{W}(\mathbf{x})\omega]\hat{V}(\omega)\phi(\omega)d\omega}{\int \exp[\hat{W}(\mathbf{x})\omega]\hat{V}(\omega)\phi(\omega)d\omega},$$

$\sigma^2(\theta|\mathbf{x})$ is approximated by

$$\hat{\sigma}^2(\theta|\mathbf{x}) = \frac{\int [\omega - \hat{E}(\theta|\mathbf{x})]^2 \exp[\hat{W}(\mathbf{x})\omega]\hat{V}(\omega)\phi(\omega)d\omega}{\int \exp[\hat{W}(\mathbf{x})\omega]\hat{V}(\omega)\phi(\omega)d\omega},$$

$H^*(\mathbf{X}|\theta)$ is approximated by

$$\hat{H}(\mathbf{X}|\theta) = -(nI)^{-1} \sum_{v=1}^{n} [p^*(\mathbf{X}_v)]^{-1} \int \hat{p}^*(\mathbf{X}_v|\omega)[\log \hat{p}^*(\mathbf{X}_v|\omega)]\phi(\omega)d\omega,$$

and $H(\theta|\mathbf{X})$ is approximated by

$$\hat{H}(\theta|\mathbf{X}) = -n^{-1}\sum_{v=1}^{n}\int \hat{f}(\omega|\mathbf{X}_v)[\log \hat{f}(\omega|\mathbf{X}_v)]d\omega.$$

In the latent-class case, $p(\mathbf{x})$ is approximated by

$$\hat{p}(\mathbf{x}) = \sum_{k=1}^{K}\hat{p}(\mathbf{x}|\tau_k)\hat{\pi}_k,$$

$f^*(\tau_k|\mathbf{x})$ is approximated by

$$\hat{f}(\tau_k|\mathbf{x}) = \hat{p}(\mathbf{x}|\tau_k)\hat{\pi}_k/\hat{p}(\mathbf{x}),$$

$E(\theta|\mathbf{x})$ is approximated by

$$\hat{E}(\theta|\mathbf{x}) = \frac{\sum_{k=1}^{K}\tau_k \exp[\hat{W}(\mathbf{x})\tau_k]\hat{V}(\tau_k)\hat{\pi}_k}{\sum_{k=1}^{K}\exp[\hat{W}(\mathbf{x})\tau_k]\hat{V}(\tau_k)\hat{\pi}_k},$$

$\sigma^2(\theta|\mathbf{x})$ is approximated by

$$\hat{\sigma}^2(\theta|\mathbf{x}) = \frac{\sum_{k=1}^{K}[\tau_k - \hat{E}(\theta|\mathbf{x})]^2 \exp[\hat{W}(\mathbf{x})\tau_k]\hat{V}(\tau_k)\hat{\pi}_k}{\sum_{k=1}^{K}\exp[\hat{W}(\mathbf{x})\tau_k]\hat{V}(\tau_k)\hat{\pi}_k},$$

$H^*(\mathbf{X}|\theta)$ is approximated by

$$\hat{H}(\mathbf{X}|\theta) = -(nI)^{-1}\sum_{v=1}^{n}[p^*(\mathbf{X}_v)]^{-1}\sum_{k=1}^{K}\hat{p}^*(\mathbf{X}_v|\tau_k)[\log \hat{p}^*(\mathbf{X}_v|\tau_k)]\hat{\pi}_k,$$

and $H(\theta|\mathbf{X})$ is approximated by

$$\hat{H}(\theta|\mathbf{X}) = -n^{-1}\sum_{i=1}^{n}\hat{f}(\tau_k|\mathbf{X}_v)[\log \hat{f}(\tau_k|\mathbf{X}_v)].$$

In all cases, $\Delta(\mathbf{X}|\theta)$ has estimate

$$\hat{\Delta}(\mathbf{X}|\theta) = \hat{H}(\mathbf{X}) - \hat{H}(\mathbf{X}|\theta),$$

so that $\Delta(\theta|\mathbf{X})$ has estimate

$$\hat{\Delta}(\theta|\mathbf{X}) = I\hat{\Delta}(\mathbf{X}|\theta).$$

One then estimates $H^*(\theta)$ by

$$\hat{H}(\theta) = \hat{H}(\theta|\mathbf{X}) + \hat{\Delta}(\theta|\mathbf{X}).$$

The estimate of $E(\sigma^2(\theta|\mathbf{X}))$ is

$$\hat{E}(\sigma^2(\theta|\mathbf{X})) = n^{-1}\sum_{v=1}^{n}\hat{\sigma}^2(\theta|\mathbf{X}_v),$$

the estimate of $E(\theta)$ is

$$\hat{E}(\theta) = n^{-1} \sum_{v=1}^{n} E(\theta|\mathbf{X}_v),$$

the estimate of $\sigma^2(\theta)$ is

$$\hat{\sigma}^2(\theta) = \hat{E}(\sigma^2(\theta|\mathbf{X})) + n^{-1} \sum_{v=1}^{n} [\hat{E}(\theta|\mathbf{X}_v) - \hat{E}(\theta)]^2,$$

and the estimate of $\rho^2(\theta|\mathbf{X})$ is

$$\hat{\rho}^2(\theta|\mathbf{X}) = 1 - \hat{E}(\sigma^2(\theta|\mathbf{X}))/\hat{\sigma}^2(\theta).$$

2 Empirical Results

Results for the Praxis data described in the introduction are summarized in Table 1. Recall that there are $I = 45$ items and $n = 8,686$ examinees. Computations employed Fortran 95 programs designed by the author for use in item-response analysis. The column with label KR is the Kuder-Richardson estimate (Kuder & Richardson, 1937) of the reliability of $\hat{W}(\mathbf{X})$ The latent-variable cases use $K = 5$ and $\tau_k = (k - 3)/2^{1/2}$, so that θ would have mean 0 and variance 1 if all latent-class probabilities π_k were equal.

To illustrate relationships between columns, it is helpful to observe that

$$\hat{\Delta}(\theta|\mathbf{X}) = 45[\hat{H}(\mathbf{X}) - \hat{H}(\mathbf{X}|\theta)].$$

For instance, in the normal 2PL case, 0.917 is $45(0.592 - 0.571)$, except for rounding error. In the computation of $\hat{\rho}^2(\theta|\mathbf{X})$, it is helpful to note that the estimated variance $\hat{\sigma}^2(\theta)$ of θ in the

Table 1

Results for 0PL, 1PL, and 2PL Models

Model	$\hat{H}(\mathbf{X})$	$\hat{H}(\mathbf{X}\|\theta)$	$\hat{\Delta}(\theta\|\mathbf{X})$	$\hat{\rho}^2(\theta\|\mathbf{X})$	KR
Latent 0PL	0.669	0.652	0.756	0.793	0.816
Normal 0PL	0.669	0.652	0.773	0.786	0.816
Latent 1PL	0.596	0.578	0.828	0.821	0.816
Normal 1PL	0.596	0.578	0.849	0.817	0.816
Latent 2PL	0.591	0.572	0.888	0.846	0.829
Normal 2PL	0.592	0.571	0.917	0.838	0.830

normal 2PL case is 1.002, a value quite close to the ideal value of 1 obtained if the model is correct and sampling error is negligible. The estimated conditional variance $\hat{\sigma}^2(\theta|\mathbf{X})$ is 0.162, so that $\hat{\rho}^2(\theta|\mathbf{X}) = 1 - 0.162/1.002$. The similarity between the columns for KR and $\hat{\rho}^2(\theta|\mathbf{X})$ is typical.

As evident from the values of $\hat{H}(\mathbf{X})$ and $\hat{H}(\mathbf{X}|\theta)$, the 0PL model is considerably less successful than is the 1PL model, and the 1PL model is a bit less successful than is the 2PL model. For the same basic model, performance of the normal and latent-class cases are quite similar. The normal cases lead to a bit larger values of $\hat{\Delta}(\theta|\mathbf{X})$ and a bit smaller values of $\hat{\rho}^2(\theta|\mathbf{X})$. The variations in reliability are rather modest, although there is some advantage with use of the 2PL model.

One method to assess the practical effects of results is to consider shortening the test and examining how the normal 2PL results compare to other reported results for 45 items. For example, with the first 30 items, $\hat{\rho}^2(\theta|\mathbf{X})$ is reduced to 0.773, a value a little lower than that reported for either 0PL model. The value of $\hat{\Delta}(\theta|\mathbf{X})$ is now 0.746, again a value a bit less than for the 0PL cases. With the first 35 items, $\hat{\rho}^2(\theta|\mathbf{X})$ is 0.797, a value slightly larger than for either 0PL model and somewhat smaller than for either 1PL model, and $\hat{\Delta}(\theta|\mathbf{X})$ is 0.804, a value somewhat larger than for either 0PL model and somewhat smaller than for either 1PL model. With the first 40 items, $\hat{\rho}^2(\theta|\mathbf{X})$ is 0.824, a value a bit larger than for either 1PL model, and $\hat{\Delta}(\theta|\mathbf{X})$ is 0.877, a value somewhat larger than for either 1PL model. Thus the results do provide some suggestion that the gain from a 2PL rather than 1PL model is sufficient to have an appreciable effect on the length of a test required to achieve comparable precision in prediction of the latent parameter θ from the observed $\mathbf{X}$.

3 Conclusions

These results are encouraging to the extent that they suggest a method of constructing an ability parameter that does not depend on model validity. The empirical results for the example considered suggest that the effects of model error are modest in terms of use of the ability parameter as a measure of examinee ability. On the other hand, the results also suggest that the old-fashioned total score gives results rather comparable in terms of reliability to results obtained by use of the theory of item responses. To the extent that these observations can be confirmed by analysis of a variety of test results, it may well be the case that the effects of model-based inference based on wrong models may often be sufficiently small not to cause significant problems in practice. This conclusion is most likely to hold for well-constructed and relatively long tests in

which Rasch or two-parameter logistic models are employed, for estimated item discrimination is likely to vary less, the effects of the latent ability distribution on the posterior ability distribution are reduced, and the model error is relatively small. One major aid is that the estimation of the ability parameter still depends on the total score or on a positively weighted sum. Because the model leads to a posterior distribution of the ability parameter given the responses that is a smooth function of the total score or weighted total, standard statistical results for functions of sums of independent variables indicate that it is indeed reasonable to anticipate that results based on estimation of the ability parameter are quite comparable to those for the total or weighted total.

It should be noted that the approach used in this paper can be considered for multivariate ability distributions as well as for univariate ability distributions and it can be applied to polytomous items. No inherent change in methodology is required.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Gilula, Z., & Haberman, S. J. (2000). Density approximation by summary statistics: An information-theoretic approach. *Scandinavian Journal of Statistics*, *27*, 521–534.

Haberman, S. J. (1979). *Analysis of qualitative data* (Vol. 2). New York: Academic Press.

Haberman, S. J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, *77*, 568–580.

Haberman, S. J. (2005). *Latent-class item response models* (ETS Research Rep. No. RR-05-00). Princeton, NJ: ETS.

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (ETS Research Rep. No. RR-06-29). Princeton, NJ: ETS.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986–1005.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.