DOCUMENT RESUME

ED 079 351 TM 002 950

AUTHOR	Lord, Frederic M.
TITLE	Estimation of Latent Ability and Item Parameters When There Are Omitted Responses.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	ETS-RB-73-37
PUB DATE	May 73
NOTE	32p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Mathematical Models; *Probability Theory; Psychometrics; *Scoring Formulas; Technical Reports; *Test Results

ABSTRACT

Omitted items cannot properly be treated as wrong when estimating ability and item parameters. A convenient method for utilizing the information provided by omissions is presented. Some theoretical and considerable empirical justification is adduced for the estimates obtained by both old and new methods. (Author)

ED 079351

TM 002 950

RESEARCH BULLETIN

RB-73-37

ESTIMATION OF LATENT ABILITY AND ITEM PARAMETERS WHEN THERE ARE OMITTED RESPONSES

Frederic M. Lord

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
May 1973

FILMED FROM BEST AVAILABLE COPY

Estimation of Latent Ability and Item Parameters When There Are Omitted Responses

Abstract

Omitted items cannot properly be treated as wrong when estimating ability and item parameters. A convenient method for utilizing the information provided by omissions is presented. Some theoretical and considerable empirical justification is adduced for the estimates obtained by both old and new methods.

Estimation of Latent Ability and Item Parameters When There Are Omitted Responses*

At the time the likelihood equations for item characteristic curve (icc) theory were written down [Lord, 1953], there seemed to be three major obstacles to practical applications:

1. Solution of the likelihood equations for data of real interest did not seem practicable from a computational point of view [Torgerson, 1958, p. 388].
2. Icc theory dealt with unspeeded tests, whereas almost all standard tests are administered with a time limit that prevents some examinees from finishing.
3. Icc theory was first developed for dichotomous items. Typical test answer sheets, however, carry at least three distinct types of examinee response: correct response, wrong response, no response ("omits"). These three types of response often receive different scoring weights (for example, 1, $-\frac{1}{4}$, and 0, respectively), but even if omitted responses are scored as wrong they can not reasonably be treated as wrong in the likelihood equations.

Solution of equations. Numerical solutions to the likelihood equations have now successfully been obtained for large data sets [Lord, 1968; Bock & Lieberman, 1970; Bock, 1972]. Even the maximum likelihood estimate of the parameter representing the icc lower asymptote, sometimes incorrectly called the chance level, can now be obtained by maximum likelihood

---

*Research reported in this paper has been supported by grant GB-32781X from National Science Foundation.

[Wingersky & Lord, 1973] whenever there is enough data to determine this part of the curve.

Speededness. Time-limit tests are, for some examinees, partly a measure of something called speed, which is quite distinct from the ability measured by the power score that would be obtained if the test were administered without time limit. Icc theory is not presently equipped to deal with the speed dimension explicitly, but the theory can still be used to analyze answer sheets obtained in timed test administrations. To do this requires the assumption that examinees answer test questions in order. For each examinee, the items following his last recorded response (this item is called the last item attempted) are ignored throughout the estimation process. (In practice, examinees answering less than a third of the n items are omitted from all analyses.) Thus examinee ability θ is estimated from his responses to items actually attempted and item parameters are estimated from the responses of examinees who attempt the item.

This does not complicate the likelihood equations or the process of solving them. A key property of icc theory is that item parameters do not depend on the group of examinees tested, within reasonable limits; and that examinee ability (θ) does not depend on the items administered, assuming that all items measure the same psychological dimension. Thus, ignoring various examinees and various items ought not to have serious effects.

It is true that a θ estimated in this way will approximate the examinee's ability under power conditions only if his responses to the items actually attempted would have been the same in the absence of a time limit. Regardless of this, a θ estimated in this way appropriately reflects the examinee's effective ability level under the timed conditions actually provided.

If an examinee does not reach an item because of the time limit, this fact contains no usable information for inferring his ability level θ under the icc model. The one-dimensional icc model considered here provides no way to make use of any relationship (ordinarily curvilinear) that may exist between speed of response and θ . The term omitted response or simply omit will hereafter refer only to items actually reached, not to items after the "last item attempted."

Omitted items. The present paper proposes and discusses a method for the effective use of the information represented by omitted responses when estimating ability and item parameters. Another method has been proposed and used by Bock [1972]. With the implementation of adequate methods for dealing with omitted responses, the application of icc theory to typical testing data is now effective and practical.

Omitted Responses

The meaning of an omitted response varies depending on the type of item and how the test is scored. It will be assumed throughout this paper that the items are multiple choice. If the score is the number of right answers, to be denoted by x , then the examinee who omits any item is acting against his own best interests. We will not consider such cases here.

The only common alternative to the number-right score is the formula score containing a penalty for wrong answers. If A is the number of alternative responses provided for the test item, the usual formula score is

$$(1) \qquad y_a = x_a - \frac{w_a}{A - 1} \quad ,$$

where w_a is the number of wrong responses given by examinee a .

We will assume hereafter that (1) is used; also that examinees wish to maximize their expected scores and that they are fully informed about their best strategy for doing this. Under these conditions, an examinee should omit an item only if he believes his chance of success on the item is no greater than $c \equiv 1/A$. On the other hand, since the item has A alternative responses, his chance of success cannot be less than c , since he can always do this well by strictly random guessing. Following this reasoning, we will assume hereafter that if an examinee were required to respond to a long series of A -choice items that he had omitted, his proportion of correct answers would be c . [Slakter (1968) presents empirical evidence that examinees omit more items than they should according to this assumption; however, his examinees were not explicitly instructed as to their best strategy.]

The ICC Model

In item characteristic curve (icc) theory, the probability that examinee a will answer item i correctly may be denoted by $P_{ia} \equiv P_i(\theta_a)$, here assumed to be an increasing function of his ability θ_a . It might seem from the preceding paragraph that $P_i(\theta_a) = c$ whenever examinee a is required to answer an item i that he previously had omitted. This cannot be correct, however. If examinee b omits the same item, we would have $P_i(\Theta_b) = c$, from which it would follow that $\Theta_a = \Theta_b$. Since two examinees who omit the same item may be at very different ability levels, it is clear that the probability c is a different kind of probability than P_{ia} .

It might seem natural to think of P_{ia} as the relative frequency of correct answers when item i is repeatedly administered to examinee a under some hypothetical conditions requiring him to forget his previous responses. This interpretation of P_{ia} is considered in detail by Meredith [1965]. We cannot use this interpretation here (nor in most practical work). Examinee a might know the answer to item $i = 1$ and so have a probability of 1 of answering it correctly. At the same time, he might be misinformed about item $i = 2$ and so have a probability of 0 of answering it correctly. At the same time, examinee b might have probability 0 of answering item 1 correctly and probability 1 of answering item 2 correctly. If the items measure the same trait, the four equations $P_1(\Theta_a) = P_2(\Theta_b) = 1$ and $P_2(\Theta_a) = P_1(\Theta_b) = 0$ are very difficult to reconcile.

P_{ia} is most simply interpreted as the probability that examinee a will give the right answer to a randomly chosen item having the same icc as item i . An alternative interpretation is that $P_i(\Theta_a)$ is the probability that item i will be answered correctly by a randomly chosen examinee of ability level $\Theta = \Theta_a$. These two interpretations will usually be assumed to hold simultaneously.

These interpretations tell us nothing about the probability that a specified examinee will answer a specified item correctly. It is this last probability that would equal c if an examinee were required to answer an item he has omitted.

The Likelihood Function

Let the response of examinee a to item i be denoted by u_{ia} . For a correct response, let $u_{ia} = 1$; for an incorrect response, let $u_{ia} = 0$. When examinee a answers a test composed of n items, the u_{ia} are assumed independent (assumption of local independence). If he does not omit any items, the likelihood function for his responses can be written

$$(2) \qquad L_a(u_{1a}, u_{2a}, \ldots, u_{na} | \theta_a) = \prod_{i=1}^{n} P_{ia}^{u_{ia}} Q_{ia}^{1-u_{ia}}$$

where $Q_{ia} \equiv 1 - P_{ia}$. If the item parameters in P_{ia} that characterize item i are known (approximately, from pretesting), then the maximum likelihood estimate $\hat{\theta}_a$ of the examinee's ability can be obtained from this likelihood function by standard procedures.

If the item parameters are unknown, they can be estimated at the same time as θ_a from the responses of many (preferably two or three thousand) examinees. In this case, the likelihood function is

$$(3) \qquad L(U | \underset{\sim}{\theta}) = \prod_{a=1}^{N} \prod_{i=1}^{n} P_{ia}^{u_{ia}} Q_{ia}^{1-u_{ia}} \quad ,$$

where U is the matrix $\|u_{ia}\|$ and $\underset{\sim}{\theta}$ is the vector $\{\theta_1, \theta_2, \ldots, \theta_N\}$. The maximum likelihood estimates of $\underset{\sim}{\theta}$ and of the item parameters can actually be obtained in practice from this likelihood function by standard procedures (somewhat surprisingly, in view of the very large number of parameters to be estimated). The practical effectiveness of this procedure is being demonstrated in work with real data (for example, Lord, 1970) despite the (presumably temporary) lack of a rigorous proof that the maximum likelihood estimates are consistent.

If the examinee omits certain items, it might seem that one could simply omit these items altogether from (2) or (3) and proceed as before. This cannot be right, however, since the fact that the examinee omitted certain items carries the important information that he did not know the answers to these items--that his chance of success was roughly only c on each. We cannot afford to ignore this information. If we did, an examinee could obtain as high a $\hat{\theta}$ as he wished, simply by omitting questions whenever he was not completely sure of the correct answer.

One way to deal with this situation would be for the psychometrician to replace each omitted response by a response drawn at random with probability of success c . After all, this is just what some examinees do in actual practice, instead of omitting items. According to the model, the likelihood function (2) or (3) will hold for data so obtained.

Although this procedure should yield consistent estimates (as $n \to \infty$), it is objectionable from two related points of view. From the examinee's point of view, it is unfair to saddle him with a possibly unfortunate set of random responses. From the statistician's point of view, the procedure degrades the data by introducing random error; it can only increase error variance, it cannot possibly be truly beneficial.

It would be desirable to replace (2) or (3) by a likelihood function that includes provision for omitted responses. Such a function, however, would depend in part on the true probability that examinee a will omit a randomly chosen item having the same item parameters as item i ($i = 1,2,\ldots,n$). This true probability would be a function, similar to $P_i(\Theta_a)$ but not the same, depending not only on Θ_a and on certain characteristics of the item, but also on a new trait of the examinee representing his willingness to omit items. Even after simplifying assumptions, there would be at least one new examinee parameter and one new item parameter to estimate, considerably complicating the already complicated and expensive estimation procedure.

New Estimation Procedure

The following estimation procedure has been used on several large sets of data, apparently with great success, as briefly indicated in the next section. The likelihood function (2) is replaced by

$$(4) \qquad L_a^*(v_{1a}, v_{2a}, \ldots, v_{na} | \Theta_a) = \prod_{i=1}^{n} P_{ia}^{v_{ia}} Q_{ia}^{1-v_{ia}}$$

where $v_{ia} = 1, 0,$ or c according to whether the response is right, wrong, or omitted, respectively. Since for the present the item parameters must be estimated at the same time as Θ_a , the responses of many (preferably two or three thousand) examinees are analyzed simultaneously, so that (4) is replaced by

$$(5) \qquad L^*(V|\underset{\sim}{\theta}) = \prod_{a=1}^{N} \prod_{i=1}^{n} P_{ia}^{v_{ia}} Q_{ia}^{1-v_{ia}} \quad ,$$

where V is the matrix $\|v_{ia}\|$. In the new procedure as actually used, we find the values of θ_a for the N examinees and the values of three item parameters for the n items that maximize (5). These values are taken as the parameter estimates desired.

Since (4) and (5) are not likelihood functions, these estimates are not maximum likelihood estimates. The estimate of θ from (4) or (5) will be denoted by θ^*. It is shown in the Appendix that in the case of (4), the θ^* converge for large n to the same values as do the $\hat{\theta}$ obtained from (2) after omits have been replaced by random responses. Moreover, if there are omits, the sampling error of θ^* for large n is less than the sampling error of the maximum likelihood estimate obtained from the degraded data.

Discussion of Assumptions

Many people prefer number-right scoring to the formula scoring considered here. Some common misinterpretations will be avoided by pointing out that certain assumptions are not made here.

1. There is no assumption here that formula scores are superior to number-right scores. If number-right scores were used, the problem considered here should not arise, since in that case examinees should not omit any items at all.

2. There is no assumption here that examinees guess at random when they do not know the answer to an item. On the contrary, many of the item characteristic curves found so far in the analysis of nationally used tests show that low-ability examinees tend to do less well on difficult items than they would have done if they had responded at random. This situation presumably arises because certain of the possible item responses have been cleverly made so attractive that low-level examinees tend to choose them in preference to the correct answer.

3. The model given here is consistent with the obvious fact that examinees use misinformation and partial information in answering items. For a few items in nationally administered tests, the icc never go below $P_i(\theta) = .30$ or $P_i(\theta) = .40$, regardless of θ level. This suggests that on some items even very low-level examinees may be able to rule out two or three of the possible item responses as incorrect.

It is assumed that the probability of a correct answer would be $c \equiv 1/A$ if an examinee were required to respond to the A-choice items he has omitted. As explained previously, this assumption is made because an examinee who wants to maximize his expected score should never omit an item if he can do better than choose among the A responses at random.

Empirical Results

Several questions can be raised about the various parameter estimates that have been discussed. In the first place, many people distrust the assumptions of the icc model, particularly the assumption that there is only one dimension θ underlying the test. The best way to resolve this question is not to try to prove that the assumptions hold for a particular set of data (they will never hold exactly), but to show that the parameter estimates obtained provide a useful and effective summary of the data, capable of predicting new sets of data not yet observed. The main purpose of this section is to show just that, insofar as possible with the limited investigations made to date.

Maximum likelihood estimates obtained from (3) are open to a further objection--there exists no rigorous proof that these estimates are consistent. A related but distinct problem is that it may seem hard to believe that several thousand parameters can really be successfully estimated simultaneously. It would be valuable to have a mathematical proof of the asymptotic properties of the estimates. Any final answer to both questions, however, must come by demonstrating the usefulness of estimates obtainable in practice from samples of reasonable size.

Estimates obtained from (4) and (5) are open to further objections. Equations (4) and (5) are not likelihood functions. No clear statistical justification has been given for choosing these functions to be maximized (there seems to be an interesting and unanswered question in statistical inference here). Some justification of the estimates obtained are

2. Relation of Estimated Ability to Test Score

Figure 1 shows for test m the relation of the ability estimates θ^*_a obtained from (5) to formula score on the total test. Under the probabilistic icc model, scatter of scores about the test characteristic curve (the regression of test score on ability) is to be expected, because of sampling fluctuations provided for in the model. The correlation ratio of formula score on $\hat{\theta}$ was .978. For test v, a correlation ratio of .982 was computed, but coarse grouping of formula scores makes this somewhat too low.

3. Predicting Test Score from Estimated Ability

For given θ_a, the expected number-right score of examinee a is

$$(6) \qquad \mathcal{E}x_a = \sum^{n_a} P_i(\theta_a)$$

where $\sum^{n_a}$ represents summation over all items actually answered. The correlation between $\sum^{n_a} P^*_i(\theta^*_a)$ and number-right score corrected for omits was obtained for tests V and m. Here $P^*_i(\theta^*_a)$ represents $P_i(\theta_a)$ with parameter estimates from (5) substituted for the unknown values. The number-right score corrected for omits is $x_a + o_a/A$, where o_a is the number of omitted items. For 60-item m, the correlation was .981; for 90-item V, .992.

These high correlations show that the estimated parameters summarize the data on the examinees' answer sheets very effectively.

2. Relation of Estimated Ability to Test Score

Figure 1 shows for test m the relation of the ability estimates θ_a^* obtained from (5) to formula score on the total test. Under the probabilistic icc model, scatter of scores about the test characteristic curve (the regression of test score on ability) is to be expected, because of sampling fluctuations provided for in the model. The correlation ratio of formula score on $\hat{\theta}$ was .978. For test v , a correlation ratio of .982 was computed, but coarse grouping of formula scores makes this somewhat too low.

3. Predicting Test Score from Estimated Ability

For given θ_a , the expected number-right score of examinee a is

$$(6) \qquad \mathcal{E}x_a = \sum^{n_a} P_i(\theta_a)$$

where $\sum^{n_a}$ represents summation over all items actually answered. The correlation between $\sum^{n_a} P_i^*(\theta_a^*)$ and number-right score corrected for omits was obtained for tests V and m . Here $P_i^*(\theta_a^*)$ represents $P_i(\theta_a)$ with parameter estimates from (5) substituted for the unknown values. The number-right score corrected for omits is $x_a + o_a/A$, where o_a is the number of omitted items. For 60-item m , the correlation was .981; for 90-item V , .992.

These high correlations show that the estimated parameters summarize the data on the examinees' answer sheets very effectively.

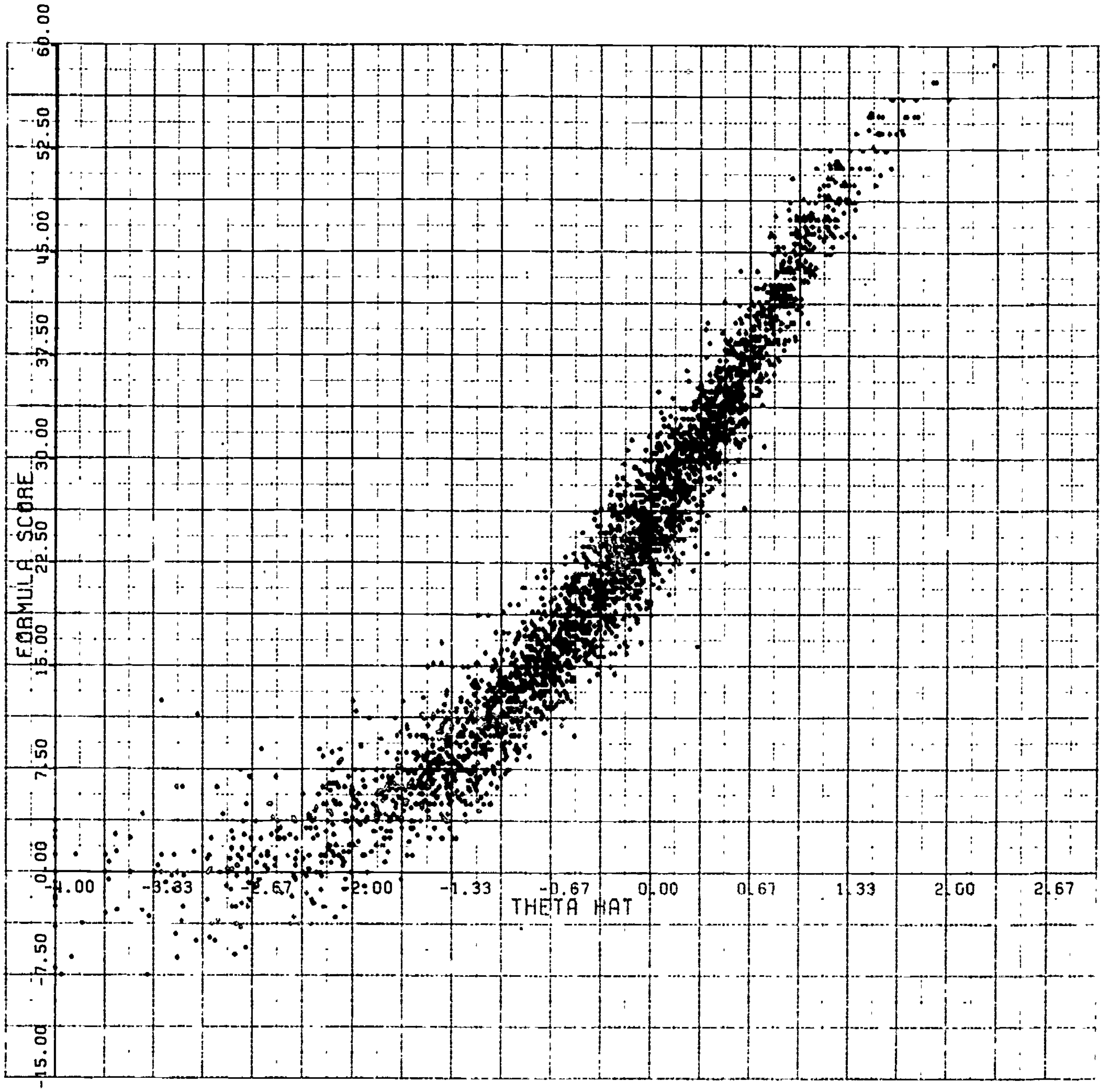

Figure 1. Relation [illegible] ability ($\hat{\theta}$) [illegible]

Again, some scatter of scores about their expected value, due to sampling fluctuations, is provided for in the model. The scatter is less than it would be if true parameters had been used instead of estimated parameters. The reason is that chance irregularities in the data are to some extent fitted in the course of the estimation process. Cross validation procedures, using a second random sample of data, could be used to eliminate this.

4. Comparing Estimates of the Distribution of Ability

The histogram in Figure 2 shows for the 1807 examinees who answered the last item in test m the frequency distribution of $\hat{\theta}$ obtained from (3). The smooth curve shows an estimate $\hat{h}(\theta)$ of the frequency distribution of θ (not $\hat{\theta}$) obtained by the method outlined below. The two distributions are obtained in very different ways, under totally different sets of assumptions, as detailed in Lord [1970], where icc were estimated by the two different methods and compared.

In order to obtain $\hat{h}(\theta)$, the frequency distribution $g(\xi)$ of true score ξ was estimated from the observed distribution of number-right scores under the compound-binomial error model [Lord, 1969]. By definition, true score is expected observed score, so by (6)

$$(7) \qquad \xi_a = \sum^{n} P_i(\theta_a) \quad .$$

When the item parameters are known, this equation defines ξ as a function of θ , or θ as a function of ξ . Thus the distribution of ability $h(\theta)$ was estimated for the group of 1807 examinees by first estimating the distribution $g(\xi)$ from the distribution of their number-right scores

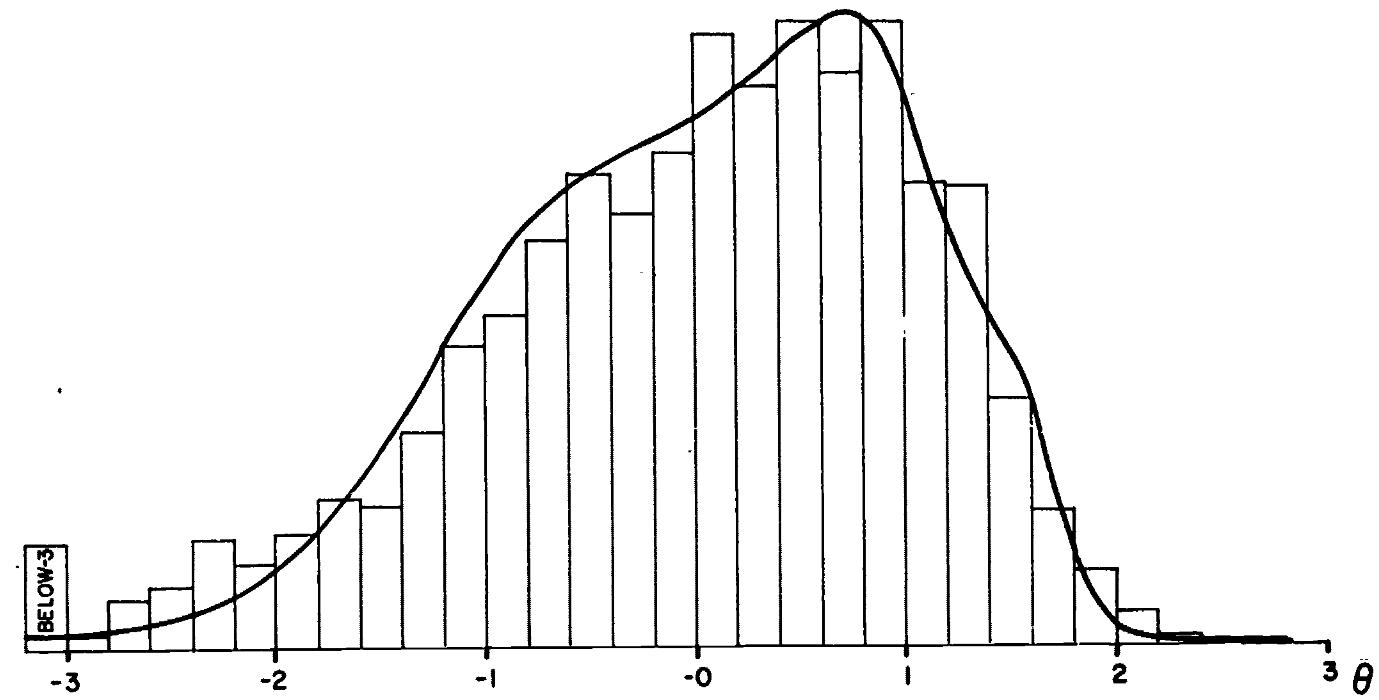

Figure 2. Distribution of Estimated $\hat{\theta}$ (histogram) and Estimated Distribution of θ (curve).

and then transforming this distribution to that of θ by the functional relationship (7) with estimated item parameters substituted for their unknown true values.

The discrepancies between the two distributions in Figure 2 occur where they should. The $\hat{\theta}$ properly show a slightly more dispersed distribution than the θ (represented by the smooth curve), since the $\hat{\theta}$ contain errors of estimation. Because m is a difficult test, these errors are quite large for low ability examinees, as discussed in an earlier section. In view of the very different assumptions made, the excellent correspondence shown in Figure 2 is strong evidence for the meaningfulness and practical usefulness of the models and estimation procedures used to obtain these results.

5. Estimated Frequency Distribution of Number-Right Scores

According to the icc model, the probability generating function [Kendall & Stuart, 1958, section 1.37] for the frequency distribution of number-right scores is

$$(8) \qquad \sum_{x=0}^{n} \phi(x)\, t^{x} \equiv \int_{-\infty}^{\infty} \prod_{i=1}^{n} [Q_i(\theta) + t\, P_i(\theta)]\, h(\theta)\, d\theta \quad .$$

Using the estimated $h(\theta)$ in Figure 2 and using estimated item parameters in $P_i(\theta)$ and $Q_i(\theta)$, the frequency distribution of number-right scores was estimated from (8). The resulting $\hat{\phi}(x)$ agreed to at least two decimal places with that obtained under the compound-binomial error model. The latter $\hat{\phi}(x)$ agrees well with the actually observed distribution of scores, the calculated chi square being near the 30-th percentile of the χ^2 distribution for 34 degrees of freedom.

6. Correlation of New Estimates and MLE

For test M, correlations were calculated between item parameter estimates obtained by the two methods (eqs. 3 and 5). For the lower asymptote (c_i) of the icc, the correlation was .990. For the discriminating power (a_i) the correlation was .995. For the difficulty parameter (b_i) the correlation was .9996. These results show that for purposes of estimating item parameters, the new estimation method yields results virtually equivalent to a maximum likelihood procedure based on the usual icc model and filled-in observations.

Tests v, V, m, and M are all difficult for low ability examinees. As a result, the ability parameter θ cannot be estimated accurately at the lowest levels--one cannot effectively distinguish between $\theta_a = -5$ and $\theta_a = -500$. This makes no practical difference as long as the problem is to predict the performance of the examinee on other tests that are about equally difficult for them. When the examinees with estimated θ's below 3.0 are omitted, the correlation between estimated θ's obtained from (3) and (5) is .997 for the remaining 2867 examinees.

7. Likelihoods

For tests M and V, a new set of data was set up for cross-validation by replacing the omits by a second set of random responses chosen independently of the first. No parameters were estimated from these cross-validation data. Instead, the estimated log L was evaluated for these data, using estimated parameters in place of the unknown true values in (3).

For test M , when the estimated parameters were MLE's obtained from the original degraded data, the estimated log L of (3) for the cross-validation data was -52348, approximately; when the estimated parameters were obtained from the original undegraded data by (5), the estimated log L of (3) for the cross-validation data was -52295. Thus the new method (5) seems to provide better estimates of the parameters of the conventional icc model (higher likelihood for the cross-validation data) than does the conventional method itself.

The test V results support the same conclusion. When the estimates were MLE's obtained from (3), log L in (3) for the cross-validation data was -58277; when the estimates were obtained by (5), log L in (3) for the cross-validation data was -58224.

Appendix A

We are given answer sheets with three kinds of responses: rights, omits, and wrongs, which will be denoted by $v_i = 1$, c, and 0, respectively. This appendix deals with the new estimation procedure for the case where there is just one examinee, the item parameters having been already determined.

Rewriting (4) with the subscript a dropped from most of the symbols, we have

$$(\text{A1}) \qquad L_a^*(v_1, v_2, \ldots, v_n | \theta) = \prod_{i=1}^{n} P_i^{v_i} Q_i^{1-v_i} \quad .$$

Our estimate of the ability of the examinee tested is θ^*, the value of θ that maximizes (A1).

If we replace all omits by randomly assigned responses, the likelihood function under the icc model is

$$(\text{A2}) \qquad L_a(u_1, u_2, \ldots, u_n | \theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \quad ,$$

where $u_i = 1$ or 0 . This equation is the same as (2) except that a has been dropped from most of the subscripts. The MLE $\hat{\theta}$ of the examinee's ability obtained from (A2) is justified by the icc model. The empirical results given in the last section suggest that θ^* may be a superior estimator to $\hat{\theta}$. It is most desirable, however, to have mathematical proofs of some of the properties of θ^* . No such proofs have been given so far. The purpose of this Appendix is to indicate some relationships between θ^* and $\hat{\theta}$ that hold for large n. It will be shown that θ^* is a consistent estimator with a sampling variance smaller than that of $\hat{\theta}$ (obtained after replacing omits by random responses).

In all that follows, we assume that θ is bounded, at least for any group of examinees that we consider testing. Another limitation is that we cannot estimate the ability of examinees who answer all items correctly or all items incorrectly. We deal only with examinees who give at least one right answer and at least one wrong answer. Since θ is bounded, the probability that an examinee will be excluded by this limitation approaches 0 for large n.

We will assume that P_i is an increasing function of θ, differentiable, with $0 \leq c_i < P_i < 1$, c_i being the lower asymptote. These assumptions are easily satisfied by all icc ordinarily used for cognitive tests. We will avoid using extremely difficult or extremely easy items, so we can assume that P_i is bounded away from 0 and 1.

Let us reorder the items so that the s omitted items are numbered $i = 1,2,\ldots,s$. Then equations (A1) and (A2) become respectively

$$\text{(A3)} \qquad L^*_a = \prod_{i=1}^{s} P_i^{c} Q_i^{1-c} \cdot \prod_{i=s+1}^{n} P_i^{u_i} Q_i^{1-u_i}$$

$$\text{(A4)} \qquad L_a = \prod_{i=1}^{s} P_i^{u_i} Q_i^{1-u_i} \cdot \prod_{i=s+1}^{n} P_i^{u_i} Q_i^{1-u_i} \quad .$$

Taking logarithms and dividing by n we can write

$$\text{(A5)} \qquad \frac{1}{n} \log L_a^* = \frac{1}{n} \sum_{i=1}^{s} c \log \frac{P_i}{Q_i} - \frac{1}{n} \sum_{i=1}^{n} \log Q_i + \frac{1}{n} \sum_{i=s+1}^{n} u_i \log \frac{P_i}{Q_i} ,$$

$$\text{(A6)} \qquad \frac{1}{n} \log L_a^* = \frac{1}{n} \log L_a - \frac{1}{n} \sum_{i=1}^{s} (u_i - c) \log \frac{P_i}{Q_i} .$$

Since the u_i are assigned at random with probability c that $u_i = 1$, the quantity Z defined by

$$\text{(A7)} \qquad Z \equiv \frac{1}{s} \sum_{i=1}^{s} (u_i - c) \log \frac{P_i}{Q_i}$$

is the average of s observations, each on a random variable having a mean of zero. If $s \to \infty$, the variance of (A7) always $\to 0$ (since Q_i is bounded away from 0). Consequently the last term in (A6) always converges in probability to zero.

Thus, for large n the likelihood function (A2) and the new function (A1) tend to the same limit. This result makes the function (A1) a plausible function to investigate, even though (A1) is itself not a likelihood function.

Likelihood Equation

The log likelihood from (A2) is $\Sigma_i u_i \log P_i + \Sigma_i (1 - u_i) \log Q_i$. Taking the derivative with respect to θ gives the result

$$\text{(A8)} \qquad \frac{d \log L_a}{d\theta} = \sum_{i=1}^{n} \left(\frac{u_i}{P_i} - \frac{1 - u_i}{Q_i} \right) P'_i \quad ,$$

where $P'_i \equiv dP_i/d\theta$. Setting this derivative equal to zero yields a familiar likelihood equation

$$\text{(A9)} \qquad \frac{d \log L_a}{d\theta} = \sum_{i=1}^{n} (u_i - P_i) \frac{P'_i}{P_i Q_i} = 0 \quad .$$

Similarly, setting $d \log L^*_a/d\theta = 0$, we obtain from (A1)

$$\text{(A10)} \qquad \frac{d \log L^*_a}{d\theta} = \sum_{i=1}^{n} (v_i - P_i) \frac{P'_i}{P_i Q_i} = 0 \ .$$

When the items are ordered with omitted items first, (A10) can be written

$$\text{(A11)} \qquad \frac{d \log L^*_a}{d\theta} = \frac{d \log L_a}{d\theta} - \sum_{i=1}^{s} (u_i - c) \frac{P'_i}{P_i Q_i} = 0 \quad .$$

Consistency

Cramér's well-known proof [1946, section 33.3] that a likelihood equation has a solution that converges in probability to the true value θ_0 as $n \to \infty$ applies with minor modifications to $\hat{\theta}$ obtained from (A9) (the cited proof only covers identically distributed variables). Cramér's approach can also be used to prove that θ^* , the solution of (A11), converges in probability to the true value θ_0 . When (A11) is divided by n and then $d \log L_a/d\theta$ expanded by a three-term Taylor's formula, we obtain after replacing θ by θ^*

$$(A12) \qquad \frac{d \log L_a^*}{d\Theta}\bigg|_{\Theta^*} = B_0 - \frac{1}{n}\sum_{i=1}^{s} (u_i - c)\frac{P_i^{*\prime}}{P_i^* Q_i^*}$$

$$+ B_1(\Theta^* - \Theta_o) + \frac{1}{2}\rho B_2(\Theta^* - \Theta_o)^2 \quad ,$$

where $P_i^* \equiv P_i(\Theta^*)$, et cetera, $|\rho| \leq 1$, and

$$B_0 = \frac{1}{n}\frac{d \log L_a}{d\Theta}\bigg|_{\Theta_o} \quad , \quad B_1 = \frac{1}{n}\frac{d^2 \log L_a}{d\Theta^2}\bigg|_{\Theta_o} \quad , \quad B_2 = \frac{1}{n}\frac{d^3 \log L_a}{d\Theta^2}\bigg|_{\Theta_1} \quad ,$$

where Θ_1 lies between Θ^* and Θ_o . The quantity $\frac{1}{n}\sum_{i=1}^{s}(u_i - c)P_i^{*\prime}/P_iQ_i$ can now be combined with the first (lowest order) term and neglected when n is large. Cramér's argument then shows that Θ^* converges to Θ_o .

Sampling Variance and Efficiency

Let $\mathcal{E}$ denote expectation over the population of items, so that $\mathcal{E}u_i = P_i^o$ and $\text{Var } u_i = P_i^o Q_i^o$, where $P_i^o = P_i(\Theta_o)$, et cetera. The asymptotic sampling variance of the maximum likelihood estimator is

$$(A13) \qquad \text{Var } \hat{\Theta} = [\mathcal{E}(\frac{d \log L_a}{d\Theta})^2 \bigg|_{\Theta=\Theta_o}]^{-1}$$

$$= [\mathcal{E}\{\sum_{i=1}^{n}(u_i - P_i^o)\frac{P_i^{o\prime}}{P_i^o Q_i^o}\}^2]^{-1}$$

$$= [\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{P_i^{o\prime}P_j^{o\prime}}{P_i^o Q_i^o P_j^o Q_j^o}\mathcal{E}(u_i - P_i^o)(u_j - P_j^o)]^{-1}$$

$$= \left(\sum_{i=1}^{n} \frac{P_i^{o\,\prime 2}}{P_i^{o2} Q_i^{o2}} \operatorname{Var} u_i \right)^{-1}$$

$$= \left(\sum_{i=1}^{n} \frac{P_i^{o\,\prime 2}}{P_i^{o} Q_i^{o}} \right)^{-1} .$$

Thus, from (A12), following Cramér,

$$\text{(A14)} \qquad \sqrt{n}(\theta^{*} - \theta_o) = \frac{\dfrac{1}{\sqrt{n}} \dfrac{d \log L_a}{d\theta}\bigg|_{\theta_o} - \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{s} (u_i - c) \frac{P_i^{*\prime}}{P_i^{*} Q_i^{*}}}{-B_1 - \frac{1}{2} \rho B_2 (\theta^{*} - \theta_o)} .$$

As n increases, $(\theta^{*} - \theta_o) \to 0$, so the second term in the denominator converges to zero. Thus, the entire denominator converges to

$$- \frac{1}{n} \mathcal{E} \frac{d^2 \log L_a}{d\theta^2}\bigg|_{\theta_o} = \frac{1}{n \operatorname{Var} \hat{\theta}} = k_o^2 \quad , \text{ say.}$$

Since $\operatorname{Var} \hat{\theta}$ is of order $1/n$, k_o^2 is independent of n to our order of approximation. If, as we suppose, $s/n \to$ some constant as $n \to \infty$, then the entire numerator is a random variable with zero mean and a finite variance, which we must now proceed to determine.

The term $\dfrac{1}{\sqrt{n}} \dfrac{d \log L_a}{d\theta}\bigg|_{\theta_o}$ is a random variable with zero mean and variance

$$\text{(A15)} \qquad \frac{1}{n}\,\mathcal{E}\left(\frac{d\log L_a}{d\Theta}\right)^2\Bigg|_{\Theta_o} = \frac{1}{n\ \mathrm{Var}\ \hat{\Theta}} = k_o^2 \quad .$$

The term $\frac{1}{\sqrt{n}}\sum_{i=1}^{s}(u_i - c)\frac{P_i^{*\prime}}{P_i^{*}Q_i^{*}}$ has zero mean. When s is fixed and large, its variance is approximately

$$\text{(A16)} \qquad \frac{1}{n}\sum_{i=1}^{s}\mathcal{E}\left[(u_i - c)^2\,\frac{P_i^{*\prime 2}}{P_i^{*2}Q_i^{*2}}\right] = \frac{1}{n}\,c(1-c)\sum_{i=1}^{s}\frac{P_i^{o\prime 2}}{P_i^{o2}Q_i^{o2}} \quad ,$$

since the u_i , $i = 1,2,\ldots, s$ are not used in computing the P_i^* . For fixed s , the covariance between the two numerator terms in (A14) is found from (A9), using the same argument, to be

$$\text{(A17)} \qquad \frac{1}{n}\,\mathcal{E}\left[\sum_{i=1}^{n}(u_i - P_i^o)\frac{P_i^{o\prime}}{P_i^oQ_i^o}\cdot\sum_{j=1}^{s}(u_j - c)\frac{P_j^{*\prime}}{P_j^*Q_j^*}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{s}\mathcal{E}\left[\{(u_i - c) + (c - P_i^o)\}\frac{P_i^{o\prime}}{P_i^oQ_i^o}(u_j - c)\frac{P_j^{*\prime}}{P_j^*Q_j^*}\right]$$

$$= \frac{1}{n}\sum_{j=1}^{s}\frac{P_j^{o\prime}}{P_j^oQ_j^o}\,\mathcal{E}\left[\frac{P_j^{*\prime}}{P_j^*Q_j^*}(u_j - c)^2\right]$$

$$= \frac{1}{n}\,c(1 - c)\sum_{i=1}^{s}\frac{P_i^{o\prime 2}}{P_i^{o2}Q_i^{o2}} \quad ,$$

approximately for large s and n .

We will need the general formulas for any random variables y , z , s :

$$(A18) \qquad \mathrm{Var}(y) = \mathcal{E}[\mathrm{Var}(y|s)] + \mathrm{Var}[\mathcal{E}(y|s)] \quad ,$$

$$(A19) \qquad \mathrm{Cov}(y,z) = \mathcal{E}[\mathrm{Cov}(y,z|s)] + \mathrm{Cov}[\mathcal{E}(y|s), \mathcal{E}(z|s)] \quad .$$

Conveniently, the second term on the right of each formula is zero for the present application. Thus, by taking the expectation of (A16) with respect to s , we obtain the unconditional asymptotic variance

$$(A20) \qquad \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{s} (u_i - c) \frac{P_i^{*\prime}}{P_i^* Q_i^*} \right] = \frac{1}{n} c(1 - c)S \quad ,$$

where

$$(A21) \qquad S \equiv \mathcal{E} \sum_{i=1}^{s} \frac{P_i^{o\prime 2}}{P_i^{o2} Q_i^{o2}} = \sum_{i=1}^{n} \frac{P_i^{o\prime 2} \omega_i}{P_i^{o2} Q_i^{o2}} \quad ,$$

a positive quantity, where ω_i is the probability that the examinee will omit an item with characteristic curve P_i^o . Similarly, the unconditional covariance corresponding to (A17) is found to be the same quantity, $\frac{1}{n} c(1 - c)S$.

From (A15) and (A20), we find the unconditional asymptotic variance of the numerator in (A14) to be

$$k_o^2 - \frac{1}{n} c(1 - c)S \quad .$$

Finally, then, the asymptotic variance of $\sqrt{n}(\theta^* - \theta^o)$ is this divided by k_o^4 or

$$\text{(A22)} \qquad \text{Var } \theta^* = \text{Var } \hat{\theta} \left[1 - c(1 - c)S \text{ Var } \hat{\theta}\right]$$

$$= \frac{1 - c(1 - c)S \Big/ \left(\sum_{i=1}^{n} P_i^{o'^2} / P_i^o Q_i^o \right)}{\sum_{i=1}^{n} P_i^{o'^2} / P_i^o Q_i^o} \; .$$

It is thus seen that the new method applied to the raw data (eqs. 4 and A1) has a smaller asymptotic sampling error than does the maximum likelihood method applied (of necessity) to the filled-in data (eqs. 2 and A2). The relative efficiency of the MLE is given by the term in brackets on the right of (A22).*

---

*The writer is indebted to Prof. Robert Jennrich for finding an error in an earlier version of this conclusion.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Bock, R. D. & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

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

Kendall, M. G. & Stuart, A. The advanced theory of statistics. Vol. 1. New York: Hafner, 1958.

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. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50.

Meredith, W. Some results based on a general stochastic model for mental tests. Psychometrika, 1965, 30, 419-440.

Slakter, M. J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968, 5, 217-221.

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

Wingersky, M. S. & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. Research Memorandum 73-2. Princeton, N.J.: Educational Testing Service, 1973.