GRADED RESPONSE MODEL OF THE LATENT TRAIT THEORY AND TAILORED TESTING

FUMIKO SAMEJIMA
University of Tennessee

INTRODUCTION

There will be no doubt about the usefulness of the latent trait theory in tailored testing, or the computer assisted adaptive individual testing. This is a pilot study for actual tailored testing, using full and partial information given by a set of graded response items. The purpose of this study is: 1) to find out how tailored testing using mostly dichotomous items can provide us with good estimates of ability compared with non-adaptive testing in which we use the full information given by the graded item responses; and 2) to find out possible branching effect of a graded item when we use one as the initial item in tailored testing. Actual data used in this study are: 1) the empirical results of paper-and-pencil tests, and 2) a hypothetical test with response patterns calibrated by the Monte Carlo method. The data analyses were partly made in such a way that we treat the data as if they were collected in actual tailored testing situations. For this reason, we call it simulated tailored testing. Terminology will be used in the same way as in Samejima's two *Psychometrika* Monographs (cf. Samejima, 1969 and 1972).

RATIONALE

The consistency of the maximum likelihood estimator when the likelihood function is given by the product of identical probability density functions or probability functions has been proved by Wald (Wald, 1949) and the proof has been shown in a simplified form by Kendall and Stuart (Kendall and Stuart, 1961, Chapter 18). In the latent trait theory, this situation corresponds to the case where all the items are equivalent, i.e., when the sets of operating characteristics of item response categories are identical for all the items, either on the dichotomous or graded response level. This, of course, is a fairly restricted case, and, in practice, we usually have to handle the sets of operating characteristics which are not identical.

The proof can easily be expanded to the case in which the probability density functions, or the probability functions, are not identical, but observations increase in number following a relatively mild restriction. Let $\xi_1, \xi_2, \ldots$ be a set of independent random variables having identical distribution with the mean μ. The strong law of large numbers, which is used in the above proof, states that for any given positive numbers ϵ and δ, there exists an N such that

$$\text{prob.}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i-\mu\right|\geqslant\epsilon\right]\leqslant\delta \text{ for every } n>N. \quad (2\text{-}1)$$

Let us define two positive integers, m and r, and consider n such that

$$n = mr, \quad (2\text{-}2)$$

where r is a fixed number, however large it may be. Let $\xi_{11}, \xi_{12}, \ldots, \xi_{1r}, \xi_{21}, \ldots, \xi_{2r}, \ldots$ be a set of independent random variables, which are classified into disjoint subsets, $A_1 = \{\xi_{11}, \xi_{12}, \ldots \xi_{1r}\}$, $A_2 = \{\xi_{21}, \xi_{22}, \ldots \xi_{2r}\}, \ldots$. Let us assume that within a subset A_j the r random variables are not necessarily identically distributed, but among the subsets we can always correspond, without overlapping, one random variable from each subset A_j ($j = 2,3, \ldots$) to each element of A_1 which has an identical distribution with that of the element of A_j with a specified mean. Let μ_k ($k = 1,2, \ldots, r$) be the mean of ξ_{1k}. If we define random variables such that

$$\zeta_j = \frac{1}{r}\sum_{k=1}^{r}\xi_{jk}, \quad (j = 1,2, \ldots) \quad (2\text{-}3)$$

then these random variables are independent and identically distributed, with the mean such that

$$E(\zeta_j) = \frac{1}{r}\sum_{k=l}^{r}\mu_r = \mu. \quad (2\text{-}4)$$

Thus the strong law of large numbers is applicable for ζ_j, if not for ξ_{jk}. Using this mild restriction, we can write

$$\lim_{n\to\infty} \text{prob.}\,[\log L_V(\hat{\theta}) < \log L_V(\theta)] = 1 \qquad (2\text{-}5)$$

where $\hat{\theta}$ is the maximum likelihood estimator of the true parameter θ, which leads to the completion of the proof of the consistency of the maximum likelihood estimator. The same restriction enables us to prove the ultimate uniqueness of the maximum likelihood estimator, the asymptotic efficiency and normality of the maximum likelihood estimator, with the asymptotic variance

$$\left\{-E\left[\frac{\partial^2}{\partial\theta^2}\log L_V(\theta)\right]\right\}^{-1}. \qquad (2\text{-}6)$$

We notice that (2-6) is the reciprocal of the test information function, $I(\theta)$. Thus if we can reasonably assume that there are at most a finite number of non-identical sets of operating characteristics and the number of items given to an examinee increases by repeating r items whose sets of operating characteristics are the same as these sets, but possibly arranged in different orders, the maximum likelihood estimator ultimately distributes normally with the true value θ as its mean and the reciprocal of the test information function as its variance. For this reason, when n is large, $I(\theta)$ can be considered as a good measure of accuracy of estimation.

Let us consider the meaning of the information function when n is relatively small. In an extreme case where $n = 1$, the test information function $I(\theta)$ equals the item information function $I_g(\theta)$. It has been shown that, as long as the model satisfies the unique maximum condition, like the normal ogive or the logistic model, the item response information function $I_{x_g}(\theta)$ is positive for the entire range of θ, except, at most, at enumerable points of θ (cf. Samejima, 1973). Under that condition, the basic function $A_{x_g}(\theta)$ such that

$$A_{x_g}(\theta) = \frac{\partial}{\partial\theta}\log P_{x_g}(\theta) \qquad (2\text{-}7)$$

is strictly decreasing in θ, and the item response information function is given by

$$I_{x_g}(\theta) = -\frac{\partial}{\partial\theta}A_{x_g}(\theta). \qquad (2\text{-}8)$$

Thus the item information function, which is given as the expectation of $I_{x_g}(\theta)$, such that

$$I_g(\theta) = E[I_{x_g}(\theta)] = \sum_{x_g=0}^{m_g} I_{x_g}(\theta)\,P_{x_g}(\theta), \qquad (2\text{-}9)$$

can be considered as the expected steepness of the basic function $A_{x_g}(\theta)$ for item g. If we consider the response pattern information function, $I_V(\theta)$, such that

$$I_V(\theta) = -\frac{\partial^2}{\partial\theta^2}\log P_V(\theta) = \sum_{x_g\in V} I_{x_g}(\theta), \qquad (2\text{-}10)$$

this is a measure of the steepness of the left hand side of the likelihood equation which is set equal to zero. The item response information function $I_{x_g}(\theta)$, therefore, is the share or contribution of each response x_g to the response pattern V of which x_g is an element, and the test information function $I(\theta)$, which can be written as

$$I(\theta) = E[I_V(\theta)] = \sum_V I_V(\theta)P_V(\theta), \qquad (2\text{-}11)$$

where $\sum_V$ means the sum over all the possible response patterns, is the expected steepness of the left hand side of the likelihood equation which is set equal to zero. Since we can interpret the steepness of the left hand side of the likelihood equation as a measure of accuracy of estimation, the test information function can be considered as a measure of accuracy of estimation even if n is relatively small. Following the same logic, the item information function $I_g(\theta)$ can be considered as the expected contribution to the accuracy of estimation by adding item g to the test. For this reason, the item information function will be given an important role in the selection of item-and-way-of-dichotomization in the present study of behavior of maximum likelihood estimates in a simulated tailored testing situation.

Suppose that we have collected testing data of n items, each of which is scored into graded categories, 0 through m_g (> 1). It has been shown that the item information function assumes much greater values for a graded item than a dichotomous item, and the problem of attenuation paradox is ameliorated for a graded item (cf. Samejima, 1969, Chapter 6). Thus it is obvious that, if we rescore each of the n items dichotomously, choosing one of the m_g category borders for dichotomization, then the accuracy of estimation of θ will be lowered. A question will be raised here: how much accuracy of estimation can we still maintain if we tailor a set of n optimal dichotomized items to an individual subject, instead of giving a set of n uniformly dichotomized items to all subjects? To find this out, we can select an initial item out of all the n items more or less arbitrarily, and treat it as if it had been presented first. If we convert the initial item to a dichotomous item by choosing one of the m_g borders for dichotomization, the examinees' item scores for that item, which range 0 through m_g, will be converted to either 0 or 1, depending on the category border used. Following the normal ogive model of the graded or dichotomous response level (cf. Samejima, 1969, Chapter 9; 1972), the first estimate, $\hat{\theta}_1$, will be

obtained. If the item score is 0, then $\hat{\theta}_1$ will be $-\infty$, if it is m_g on the graded response level or 1 on the dichotomous response level, then $\hat{\theta}_1$ will be ∞, and, otherwise, it will be a finite value. When $\hat{\theta}_1$ is negative infinity, the next item and the way of dichotomization will be chosen by searching the least value of b_{x_g} among those of the remaining (n−1) items, and, when $\hat{\theta}_1$ is positive infinity, the greatest b_{x_g} is searched and used. When $\hat{\theta}_1$ is a finite value, then the item and border which make the item information function for the dichotomized item maximum at $\theta = \hat{\theta}_1$ is chosen and treated as the second presentation. In this way, the second estimate, $\hat{\theta}_2$, will be obtained, and the process will be repeated until we get the nth estimate, $\hat{\theta}_n$.

This simulated tailored testing situation is different from the actual tailored testing situation, in the sense that the selection is more limited in later presentations of items. In the ordinary case, we start with a large set of dichotomous test items, and the number of items is reduced by one after each tailored presentation. In the present simulated tailored testing situation, however, the number of items is reduced by m_g, after the presentation of item g, and at the last presentation selection is made only out of m_h possibilities, where h is the remaining item. This will make the estimation more inefficient in later processes. and should be kept in mind when observations are made for the results of the data analysis.

EMPIRICAL DATA AND THEIR ANALYSIS

A test of 18 items was constructed for research purposes, each of which is to be scored in a graded way. It consists of two subtests, figural (FGR) and numerical (NMB), the former having ten items and the latter having eight items. The initial instructions for each subtest, and also a hypothetical NMB item, which was made for illustrative purposes are shown in Appendix A.

The test was administered to 446 subjects, mostly college and summer school students in the United States and Canada, in March through July, 1974, to get the complete data of 406 subjects. In some sessions FGR was presented first, and in some others NMB was presented first. Each session required approximately 90 minutes, including initial instructions and five minutes' break between the two subjects. The number of subjects in each session varied from one to 36, but in many cases it was less than ten. A time limit is set for each item, and is between 2 and 6 minutes, except for the last NMB item for which it is 13 minutes. When there is one more minute left for each item, the instructor calls, "One more minute to go." The full item score, m_g, is 3 for each of the FGR items and also for each of the first seven NMB items, and it is 7 for the eighth NMB item. For the FGR items, 1 is given for the completion of A and B, 2 for that of A through D, and 3 for that of A through E (cf. Appendix A). For the first seven NMB items, the score is given in accordance with the number of correct answers in each item, and for the last item the score is given in a similar way as it is for a FGR item.

It turned out that the tenth item in FGR was too difficult for most subjects, and it was excluded in the analysis of the data, to leave nine items for the subtest FGR. It also turned out that frequencies for some item score categories were too small, so suitable recategorizations were made to leave three item score categories for items 4, 6, 7 and 8 in FGR, two for item 9 in FGR, and five for item 8 in NMB, making every frequency, at least, as large as 18. For the 17 item variables, which are assumed behind the item scores, the multivariate normality was assumed, and the polychoric correlation coefficient (cf. Tallis, 1962) was computed for each pair of the item variables, using Lieberman's program (Lieberman, 1969). The principal factor solution was applied for the resulting correlation matrix using the SPSS factor analysis program with iteratively estimated communalities, to obtain eigenvalues: 5.859, 1.757, 0.902, 0.745, 0.578, etc., which prove the existence of a strongly dominating first principal factor and a moderately dominating second factor. Several different factor rotations were made, both orthogonal and oblique, for these two factors, and the results uniformly showed the two clusters, one for each of the two subsets of items, i.e., figural and numerical. Table 1 shows the results of both varimax and quartimax rotations, along with the original factor loadings for the two principal factors. For this reason, each subset of items, i.e., F1 through F9, for FGR or N1 through N8 for NMB, was analyzed separately, and the first principal factor for the figural set of items, whose eigenvalue turned out to be 3.029 or 60.2% of the total sum of communalities, was named the figural ability, and the first principal factor for the numerical set, whose eigenvalue was 4.132 or 79.5% of the total communalities, was named the numerical ability. The item parameters for the operating characteristics, which follow the normal ogive model on the graded response level (cf. Samejima, 1969 & 1972), were calculated, using the formulas:

$$a_g = \rho_g / [1 - \rho_g^2]^{1/2} \qquad (3\text{-}1)$$

and

$$b_{x_g} = \gamma_{x_g} / \rho_g \qquad \text{for } x_g = 1, 2, \ldots, m_g \ ; \qquad (3\text{-}2)$$

where ρ_g is the factor loading of item g and γ_{x_g} is the normal deviate corresponding to the proportion of the subjects who got the item score x_g or greater. These

TABLE 1

Factor Loading Matrices of the Seventeen Items for the First Two Common Factors for the Original Principal Factors, After They Were Rotated Using Varimax and Quartimax Rotations.

Item	Without Rotation		Varimax Rotation		Quartimax Rotation	
	First Factor	Second Factor	First Factor	Second Factor	First Factor	Second Factor
F1	.485	.371	.106	.601	.611	.005
F2	.612	.455	.143	.749	.762	.017
F3	.577	.386	.163	.675	.692	.050
F4	.424	.154	.207	.400	.429	.139
F5	.432	.286	.125	.503	.516	.040
F6	.433	.321	.102	.529	.539	.013
F7	.358	.174	.146	.370	.389	.083
F8	.381	.274	.113	.440	.452	.039
F9	.502	.106	.298	.418	.461	.225
N1	.683	-.344	.736	.208	.326	.691
N2	.750	-.165	.664	.386	.490	.591
N3	.580	-.346	.662	.138	.245	.630
N4	.776	-.193	.702	.383	.493	.630
N5	.524	-.410	.663	.052	.160	.645
N6	.581	-.396	.696	.102	.215	.669
N7	.826	-.133	.698	.461	.570	.613
N8	.537	.086	.337	.426	.476	.262

parameter values are presented as Tables 2 and 3 for the figural and the numerical abilities respectively.

Since there is no way of knowing each examinee's true ability score, the maximum likelihood estimate, $\hat{\theta}$, was obtained from his response pattern of graded item scores, and was treated as the best possible estimate of his true ability score. Also the test information function, which is given by Equation 2-11, was calculated for each subtest, and it turned out that the subtest NMB is far more informative than the subtest FGR. Figure 1 presents the test information function of the subtest NMB. Taking the interval, $[-0.1, 1.0]$, in which the values of the test information function are no less than 7, we let the computer search the best possible way of dichotomization of each item, to make the test information as large as possible for this interval, and the resulting test information function is drawn by a dashed line in Figure 1. A similar trial was made for the least informative way of dichotomization, and the resulting test information function is shown by a dotted line in the same figure. Selecting all the subjects whose $\hat{\theta}$ are located in the above interval, the maximum likelihood estimate was calculated for each of these 138 subjects, using both the

TABLE 2

Item Parameters For the Subtest FGR

Item g	Discrimination Index a_g	Difficulty Indices b_{x_g} $x_g = 1$	$x_g = 2$	$x_g = 3$
1	0.8972	-1.0042	-0.3356	0.0833
2	1.3196	-0.7468	-0.3532	-0.0465
3	1.0160	-1.2464	-0.5137	0.1476
4	0.5775	-0.7984	0.1730	
5	0.5940	-1.1081	0.7169	0.9554
6	0.6558	-0.0337	3.1045	
7	0.4293	0.4722	3.2345	
8	0.5644	-0.7988	2.5679	
9	0.5483	2.0052		

TABLE 3

Item Parameters For the Subtest NMB

Item g	Discrimination Index a_g	Difficulty Indices b_{x_g}			
		$x_g = 1$	$x_g = 2$	$x_g = 3$	$x_g = 4$
1	1.18738	-0.58387	0.02422	0.69302	
2	1.27938	0.91100	1.21130	1.69291	
3	0.90123	-1.97011	-1.61105	-0.87804	
4	1.44248	0.06765	0.32693	0.84445	
5	0.80989	-0.99294	-0.15721	1.00489	
6	0.93783	-0.48721	0.47768	1.71261	
7	1.58894	0.02918	0.36308	0.72073	
8	0.53530	0.14401	0.52872	1.90170	2.89123

most informative and the least informative ways of dichotomization. Figure 2 shows the sets of these estimates plotted against $\hat{\theta}$. We can see a substantial difference between the two scatter diagrams.

A question will be raised here: what will the scatter diagram be if we tailor the way of dichotomization for each individual subject? To answer this, a program was written to treat the data as if these eight items had been presented in tailored testing selecting both item and way of dichotomization, as was described at the end of the preceding section. Using the most informative dichotomized item, N7 with the category border 2, the least informative dichotomized item, N3 with the border 1, and a medium informative item, N1 with the border 2, the resulting scatter diagrams are shown in Figure 3. We can see that in all these cases extremely scattered points are rare, com-

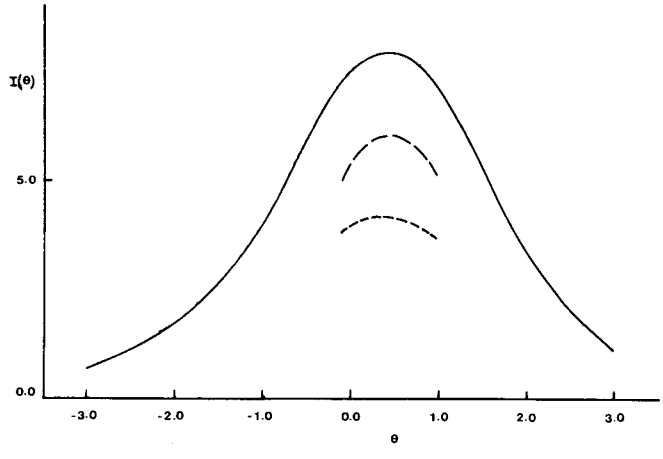


Figure 1. Test information functions for the subtest NMB, when the graded scoring strategy is taken (————), when the most informative dichotomous scoring strategy is taken for the interval [-0.1, 1.0] (– – – – –), and when the least informative dichotomous scoring strategy is taken for the interval [-0.1, 1.0] (- - - -).

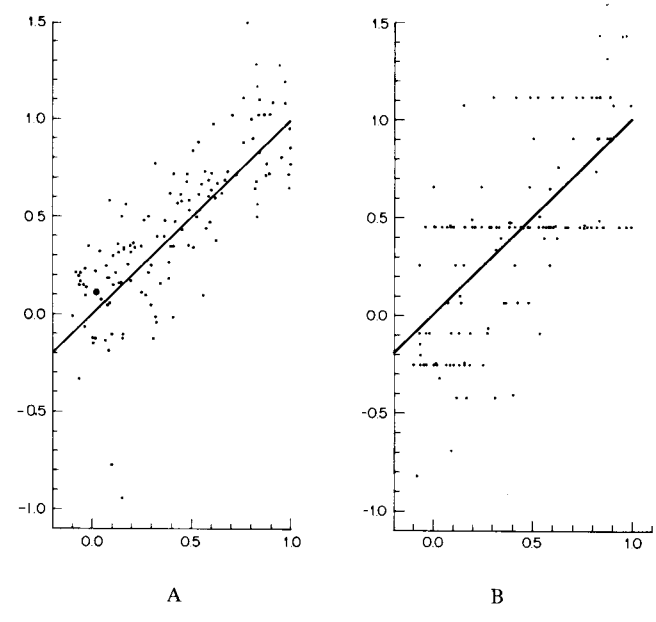


Figure 2. Maximum likelihood estimates obtained by dichotomizing NMB items for the interval [-0.1, 1.0], plotted against $\hat{\theta}$, those obtained from the original response patterns of graded item scores for the 138 subjects whose $\hat{\theta}$ are in the interval [-0.1, 1.0]. A. Using the most informative way of dichotomization, B. Using the least informative way of dichotomization.

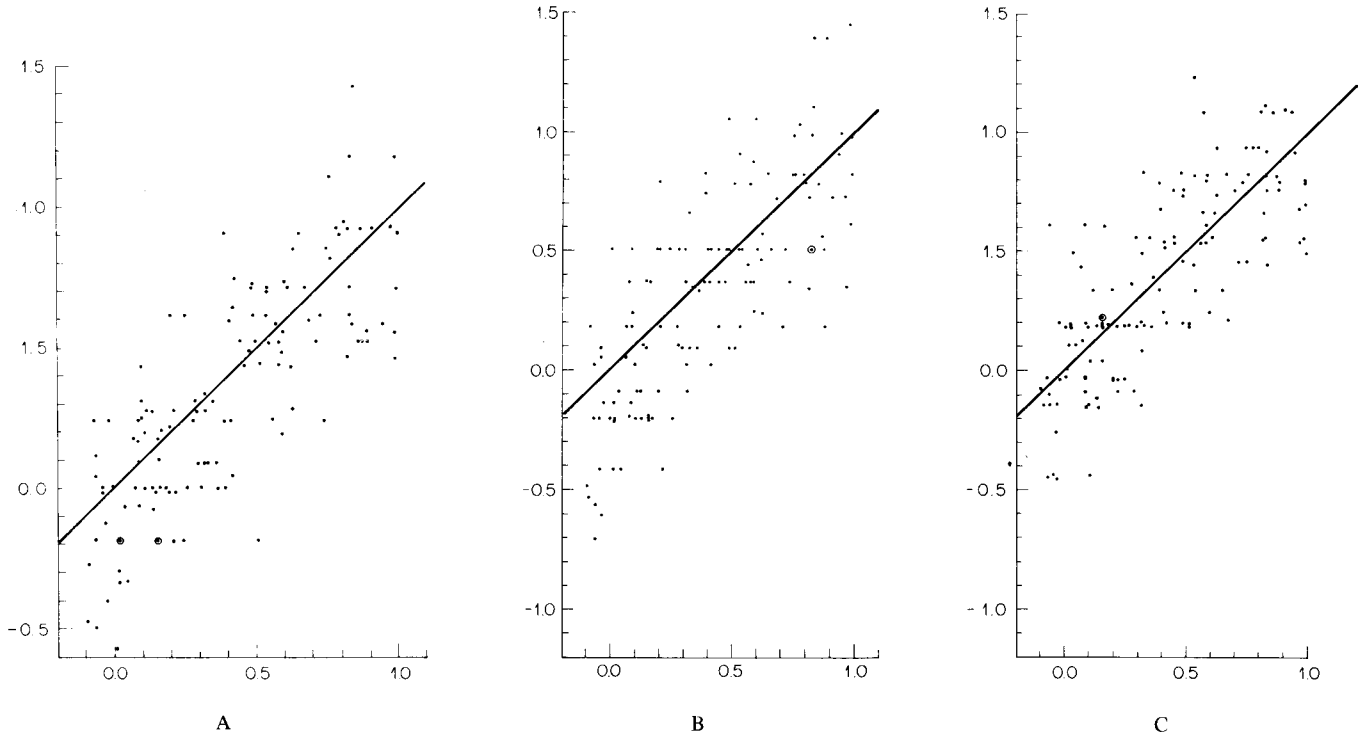


Figure 3. Maximum likelihood estimates obtained by simulated tailored testing plotted against $\hat{\theta}$, those obtained from the original response patterns of graded item scores for the 138 subjects whose $\hat{\theta}$ are in the interval [–0.1, 1.0]: A. Using the most informative dichotomized items, N7 with the category border 2, as the initial item, B. Using the least informative dichotomized item N3 with the category border 1 as the initial item, C. Using a dichotomized item of medium information, N1 with the category border 2, as the initial item.

pared with Figure 2A, i.e., the case of the most informative dichotomization for the group of these 138 subjects to say nothing about the comparison with Figure 2B. This can be interpreted as a benefit obtained by tailoring an individual test for each examinee.

A second question will be raised here: is there any substantial gain if we use a graded test item, instead of a dichotomous one, as the initial item in tailored testing? Since the number of items is as small as eight, it will be of benefit if the use of a graded item gives a substantial branching effect at the beginning of tailored testing. To find this out, using the most informative and the second most informative graded items, N7 and N4, as the initial items respectively, the same simulated tailored testing procedure was applied to obtain the maximum likelihood estimate for each individual subject. The results are shown as Figure 4. To observe the possible branching effect, in the first case the total 138 subjects were divided into two groups, one consisting of the subjects whose graded score for N7 are either 3 or 0, i.e., best or worst, and the other consisting of those who obtained either 2 or 1, i.e., intermediate scores. We can see an obvious branching effect by comparing Figures 4A and 4B.

Similar analysis was made for the other subtest, FGR and the results are presented as Appendix A. Since the maximum test information for FGR is a little more than 4 compared with that of NMB which is almost 8, there is a general tendency that diagrams are more scattered, but, other than that, similar tendencies as in NMB were observed. The interval of ability taken for these observations was [–0.8, 0.1]; there are 123 subjects whose $\hat{\theta}$ are in this interval, and the test information function for this interval is greater than 4, with an approximate maximum of 4.251 at $\theta = -0.3$. The initial items used for the simulated tailored testing are: F2 with the category border 2 (most informative), F6 with the category border 2 (least informative), F3 with the category border 3 (medium), F2 (most informative graded) and F3 (second most informative graded).

Figure 5 presents two examples to illustrate how the maximum likelihood estimate converges in the simulated tailored testing, for NMB, using the five different initial items which were described in a previous paragraph. It may be suggested that the number of items, eight, is not sufficient for all the cases. It should be recalled, however, that in the present study the selection of item-and-way-of-

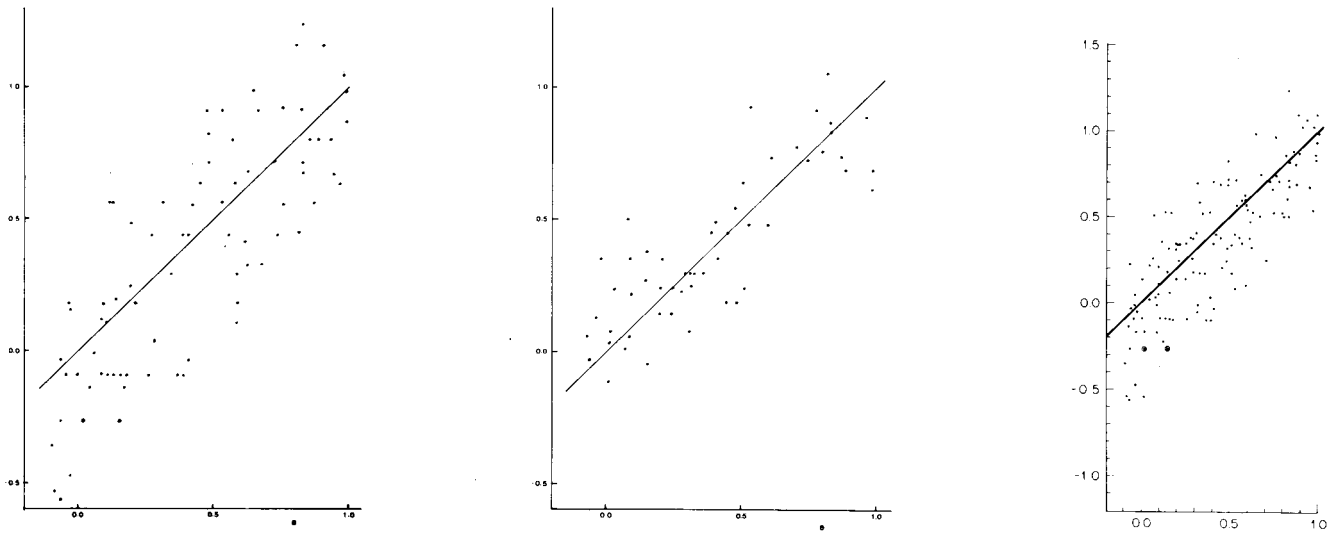


Figure 4. Maximum likelihood estimates obtained by simulated tailored testing plotted against c, those obtained from the original response patterns of graded item scores, for the subjects whose $\hat{\theta}$ are in the interval [–0.1, 1.0] : A. Using the most informative graded item, N7, as the initial item, for subjects whose item scores for N7 are extreme, i.e., either 0 or 3, B. Using the most informative graded item, N7, as the initial item, for subjects whose item scores for N7 are intermediate, i.e., either 1 or 2, C. Using the second most informative graded item, N4, as the initial item.

dichotomization is more and more limited in later presentations of items. And yet each dichotomized response pattern as a whole is a selection out of the 8,748 possibilities.

MONTE CARLO DATA AND THEIR ANALYSIS

To make further observations in the present simulated tailored testing, a hypothetical test of 24 items was used,

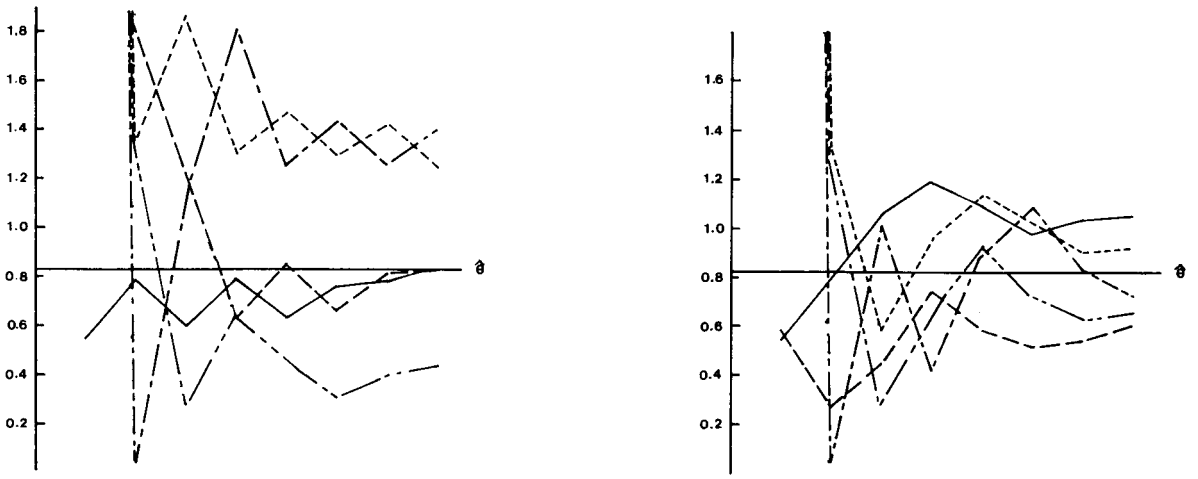


Figure 5. Two examples to show how the maximum likelihood estimates converge in the simulated tailored testing. Initial items are: N7, most informative graded item (————); N4, second most informative graded item (– – –); N7-2, most informative dichotomized item (- - - -); N1-2, medium informative dichotomized item (- • • -); and N3-1, least informative dichotomized item (- • -).

following the normal ogive model of graded response level. The item parameters were given within the range of those of NMB, so that this hypothetical test can be considered as an expansion of NMB in a rough sense of the word. Table 4 presents the item parameters of these twenty-four hypothetical items, which have uniformly four item score categories each. The test information function was obtained following the formula (2-11), and is presented as Table 5. As we can see from this table, this hypothetical test is most informative around $\theta = -0.3$. For this reason, one hundred response patterns for these twenty-four test items were calibrated by Monte Carlo method on this level of ability, and were used as those of one hundred hypothetical subjects. Figure 6 presents the cumulative frequency ratio of $\hat{\theta}$ for these response patterns, in comparison with the normal distribution function with $\mu = -0.3$ and $\sigma = 0.2128$, i.e., $1/\sqrt{22.081}$. We can see that these two curves are close, and this indicates that the maximum likelihood estimate with these parameter values already distributes almost normally for the 24 items. As before, the most informative and least informative dichotomizations of the items were searched, and the resulting maximum likelihood estimates were computed for each of these one hundred hypothetical subjects. Figures 7A and 7B present the cumulative frequency ratios of these estimates together with the normal distribution functions with $\mu = -0.3$ and the values of the standard deviation obtained by $1/\sqrt{f(-0.3)}$, which turned out to be 0.2407 and 0.3685 respectively. Since in the present situation the ability level is fixed at -0.3, the difference between the two standard deviations, 0.2128 and 0.2407, should be interpreted as the minimized reduction caused by adopting the dichotomous scoring strategy, and the one between 0.2407 and 0.3685 should be attributed to the two different ways of dichotomization. It is also noticed that the discrepancies between the normal curve and the cumulative frequency ratio are more conspicuous in these two dichotomized cases compared with Figure 6.

Figure 8 shows the same cumulative frequency ratios compared with $N(-0.3, 0.2128)$, for the maximum likelihood estimates obtained by the simulated tailored testing, with the five different initial items: (23-2), the most informative dichotomous; (3-3), the least informative dichotomous; (14-3), a medium informative dichotomous; (24), the most informative graded; and (23), the second most informative graded; respectively. The mean square errors for these five cases are 0.064, 0.068, 0.055, 0.056 and 0.058 respectively. If we take the square roots of these values, they are 0.253, 0.260, 0.234, 0.236 and 0.240, which are comparable to 0.2407, i.e., $1/\sqrt{f(-0.3)}$ for the result of the most informative dichotomization case. This is understandable because in that case the dichotomization was, indeed, tailored for the level of $\theta = -0.3$. To find out about the branching effect of the initial graded items, four more cases were added using four different dichotomized initial items of various information levels, and the results were arranged in Table 6 in the order of information levels

TABLE 4

Item Parameters of 24 Hypothetical Test Items

Item g	Discrimination Index a_g	Difficulty Indices b_{x_g}		
		$x_g = 1$	$x_g = 2$	$x_g = 3$
1	0.50000	-0.70000	-0.50000	0.20000
2	0.50000	-2.00000	-0.80000	-0.20000
3	0.60000	0.30000	0.80000	2.10000
4	0.60000	0.0	0.40000	1.30000
5	0.70000	-1.30000	-0.20000	0.40000
6	0.70000	0.20000	0.90000	2.00000
7	0.80000	-0.50000	0.80000	1.90000
8	0.80000	-1.10000	-0.90000	-0.10000
9	0.90000	-0.20000	0.40000	0.60000
10	0.90000	-1.60000	-1.00000	0.20000
11	1.00000	-1.80000	-1.10000	-0.60000
12	1.00000	0.10000	1.40000	1.60000
13	1.10000	-0.10000	0.80000	1.10000
14	1.10000	-1.00000	-0.50000	0.0
15	1.20000	-1.20000	-0.20000	0.80000
16	1.20000	-1.70000	-0.80000	-0.50000
17	1.30000	-0.30000	0.50000	1.40000
18	1.30000	-0.60000	0.40000	0.80000
19	1.40000	-0.90000	0.30000	1.10000
20	1.40000	-0.40000	-0.10000	0.60000
21	1.50000	-1.90000	-1.60000	-1.20000
22	1.50000	-1.50000	-0.40000	0.90000
23	1.60000	-0.80000	-0.40000	0.80000
24	1.60000	-1.40000	-0.60000	0.40000

of initial items. We can see from this table that, with the exception of (14-3), the values of the mean square errors are greater for the cases in which we used dichotomized items as the initial item, than those for the cases in which graded items were used, although the differences are small. To make a more detailed observation, two cases, in which (24) and (14-3) were used as the initial item respectively, were picked up, and these values were calculated for the maximum likelihood estimates when 4, 6, 8, 12, 16, 20 and 24 items were used respectively in the simulated tailored testing. The result is presented as Figure 9, in the form of the comparison of the corresponding square roots of the mean square errors. We can see that the branching effect is conspicuous for the cases of fewer items, namely, 4, 6 and 8, and disappears with the addition of more items. This can be interpreted that when we add more items the effect of the initial item becomes negligibly small. Note, however, that in the present simulated tailored testing situation the selection of item-and-way-of-dichotomization becomes more and more limited in later presentation of items.

TABLE 5

Test Information Function of the Hypothetical Test of 24 Grade Items

Ability θ	Information Function $I(\theta)$
-1.5	16.317
-1.4	17.250
-1.3	18.119
-1.2	18.915
-1.1	19.628
-1.0	20.252
-0.9	20.784
-0.8	21.220
-0.7	21.562
-0.6	21.813
-0.5	21.979
-0.4	22.065
-0.3	22.081
-0.2	22.034
-0.1	21.930
0.0	21.776
0.1	21.574
0.2	21.326
0.3	21.030
0.4	20.681
0.5	20.273
0.6	19.800
0.7	19.256
0.8	18.636
0.9	17.938
1.0	17.164
1.1	16.318
1.2	15.409
1.3	14.449
1.4	13.452
1.5	12.435

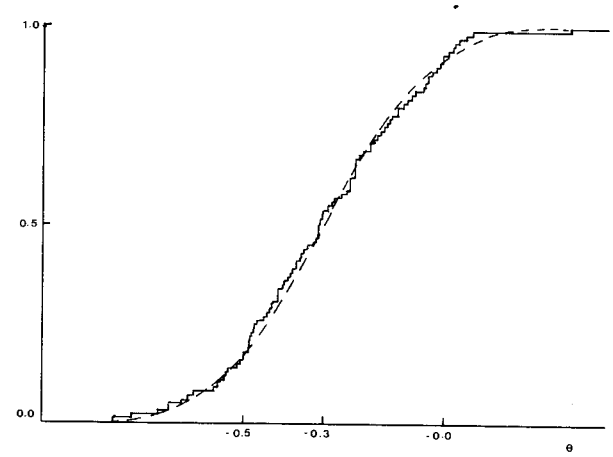


Figure 6. Cumulative frequency ratio of maximum likelihood estimates obtained from the original response patterns of graded item scores for the 100 hypothetical subjects (————) and the normal distribution function (---) with the parameters $\mu = -0.3$ and $\sigma = 0.2128$

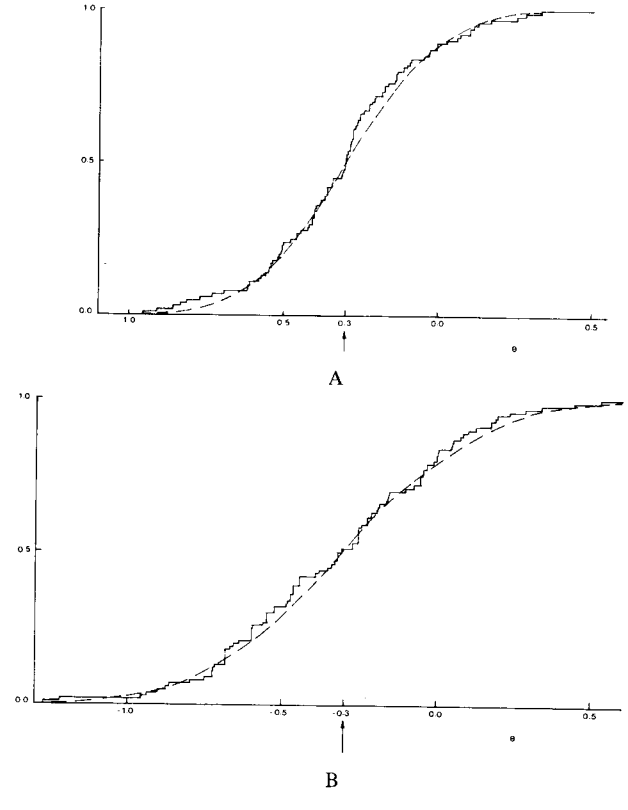


Figure 7. Cumulative frequency ratio of maximum likelihood estimates obtained from converted response patterns: A. Using most informative dichotomization of items at $\theta = -0.3$, for the 100 hypothesized subjects (————) and the normal distribution with the parameters $\mu = -0.3$ and $\sigma = 0.2407$ (– – –), B. Using least informative dichotomization of items at $\theta = -0.3$ for the 100 hypothetical subjects (————) and the normal distribution function with the parameters $\mu = -0.3$ and $\sigma = 0.3685$ (– – –).

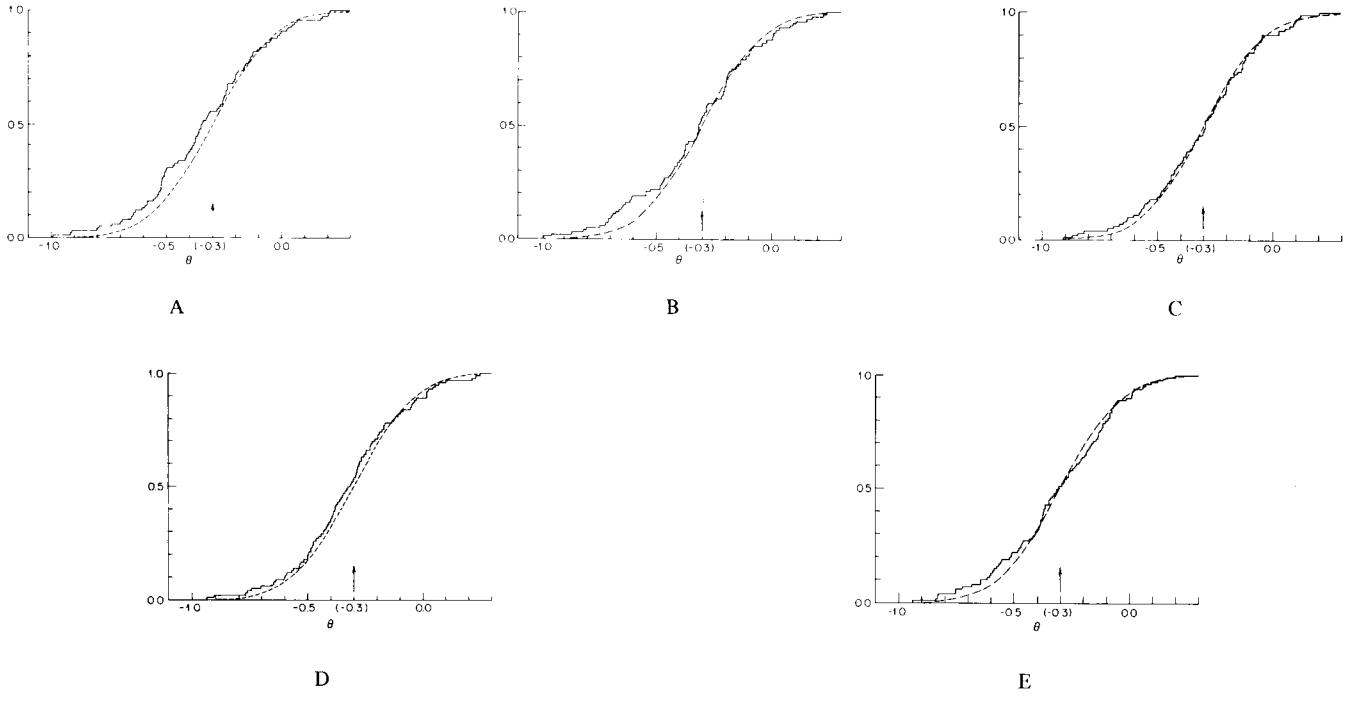


Figure 8. Cumulative frequency ratio of maximum likelihood estimates obtained by simulated tailored testing, for the 100 hypothetical subjects (– – –) and the normal distribution with the parameters $\mu = -0.3$ and $\sigma = 0.2128$ (– – –): A. with the most informative dichotomized item (23-2) as the initial item, B. with the least informative dichotomized item (3-3), as the initial item, C. with a medium informative dichotomized item (14-3) as the initial item, D. with the most informative graded item (24) as the initial item, E. with the second most informative graded item (23) as the initial item.

TABLE 6

Mean Square Errors and Other Indices for the Variability of the Maximum Likelihood Estimates in the Simulated Tailored Testing Using Different Initial Items in NMB.

	Initial Item	$I_g(-0.3)$	Mean Square Error	$\sqrt{MSE}$	1/MSE
Dichotomous	3 - 3	0.104	0.068	0.260	14.767
	5 - 1	0.260	0.069	0.263	14.430
	10 - 3	0.479	0.060	0.245	16.723
	14 - 3	0.740	0.055	0.234	18.281
	18 - 1	1.018	0.066	0.258	15.051
	23 - 1	1.287	0.063	0.250	15.938
	23 - 2	1.615	0.064	0.253	15.580
Graded	23	2.074	0.058	0.240	17.332
	24	2.127	0.056	0.236	17.980

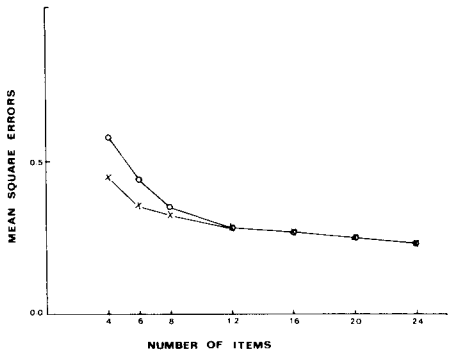


Figure 9. Comparison of the square roots of the mean square errors of the maximum likelihood estimates in simulated tailored testing with the graded item (24), plotted with x and the dichotomized item (14-3), plotted with o, as the initial item, calculated for 4, 6, 8, 12, 16, 20, and 24 items.

DISCUSSION AND CONCLUSION

Through the observations of two types of data, it has been made clear that tailored testing, in which we use dichotomous test items only, can provide us with much more accurate estimation of ability than non-adaptive testing, and that accuracy is almost comparable with that of graded response level. We also have observed that the branching effect by using a graded item as the initial item is conspicuous when we use a relatively small number of items. When the number of items increases in tailored testing, however, the effect of the initial branching, or the amount of information given by the initial item, seems to have a less effect on the final estimation. On this point, we need a further study by using a larger number of items in the original set of items, and also an item with more score categories as the initial item.

REFERENCES

Kendall, M. G. and Stuart, A. *The advanced theory of statistics.* Vol. 2. London: Griffin, 1961.

Lieberman, M. Calculation of a polychoric correlation coefficient. *Paper presented at the Psychometric Society spring meeting,* 1969, Educational Testing Service, Princeton, New Jersey.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* 1969, No. 17.

Samejima, F. A general model for free-response data. *Psychometrika Monograph,* 1972, No. 18.

Samejima, F. A comment on Birnbaum's three parameter logistic model. *Psychometrika,* 1973, 38, 221-233.

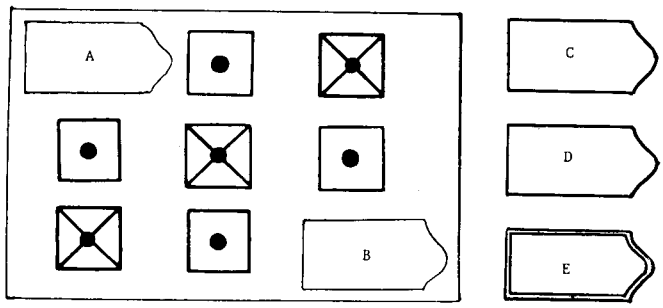
Tallis, L. R. The maximum likelihood estimation of correlation from contingency tables. *Biometrika,* 1962, 18, 342-353.

Wald, A. *Selected papers in statistics and probability by Abraham Wald.* (t. W. Anderson, et al, Ed.) Stanford University Press, 1957.

APPENDIX A

1. INSTRUCTIONS FOR THE FIGURAL SUBTEST

There are 10 items in this part of the test. In each item, nine figures are arranged in three rows and three columns, two of which are missing, as shown below. These figures are arranged according to some rule, and you must find that rule by observing the seven figures shown in the array.



Below this array, twelve figures are given, and you are to choose the right figures for the missing ones in the above array, A and B.

Next, we add an additional column as shown above. You are to choose the right figures for C and D out of the same twelve choices.

After you have followed the above two steps, then you are to draw the right figure for E in the additional column. This figure may or may not be one of the twelve choices.

Don't turn the page until you are told to do so by the instructor.

2. INSTRUCTIONS FOR THE NUMERICAL SUBTEST

There are 8 items in this part of the test. In each item, a specific rule is given, and you are to read the instruction carefully so that you will understand and be able to handle the rule. They are numerical items, and in all of them you must use calculations.

In each item, be sure that you understand the rule correctly. If you have time, check the calculations, and be sure that the (positive or negative) sign attached to your answer to each problem is a correct one. Try to solve each problem correctly and as quickly as possible.

Once you have started a calculation, continue the calculation until you get the answer. Don't leave it unfinished and start another.

Are there any questions?

Don't turn the page until you are told to do so by the instructor.

3. ITEM 1, NUMERICAL SUBTEST

The following square array of numbers is named E.

$$E = \begin{Vmatrix} 1 & 2 \\ 3 & 4 \end{Vmatrix}$$

The first column of E, $\begin{Vmatrix} 1 \\ 3 \end{Vmatrix}$, is called e_1, and its second column, $\begin{Vmatrix} 2 \\ 4 \end{Vmatrix}$ is called e_2.

Each number in a column is called an *element*. In the above example, 1 and 3 are elements of the column e_1, and 2 and 4 are elements of the column e_2.

The operator Ω indicates that you should subtract from each element of the column which comes next to the operator the corresponding element of the column which follows, square the resulting value, and then multiply all the results.

Example: $\Omega e_1\ e_2 = (1 - 2)^2 \times (3 - 4)^2 = 1$

Consider the above example(s), and *be sure that you understand the operation.*

Following this rule, compute each of the three numbers shown on the next page for the square array A, which is given below.

$$A = \begin{Vmatrix} 3 & 5 & -2 \\ -4 & 9 & -7 \\ -6 & -1 & 8 \end{Vmatrix}$$

(i) $\Omega a_1\ a_2 =$

(ii) $\Omega a_2\ a_3 =$

(iii) $\Omega a_1\ a_3 =$

If you have already finished the above, *confirm that you have used the operation correctly.* Also check the calculations, and be sure that the (positive or negative) sign attached to your answer to each problem is correct.

Don't turn the page until you are told to do so by the instructor.

APPENDIX B

Seven Figures for the Subtest FGR, Corresponding to Figures 2 through 9 for the Subtest NMB. Intial Items Used for Simulated Tailored Testing Are: F2-2 for Figure B3, F6-2 for Figure B4, F3-3 for Figure B5, F2 for Figure B6, Which Corresponds to the Combination of Figures 7 and 8 for NMB, and F3 for Figure B9. These Values Are Plotted for the 123 Subjects Whose $\hat{\theta}$ Are in the Interval [−0.8, 0.1].

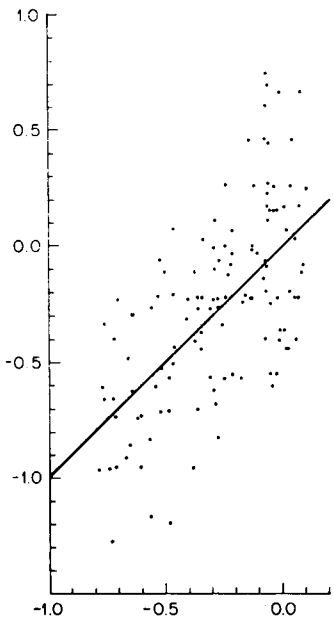


Figure B1

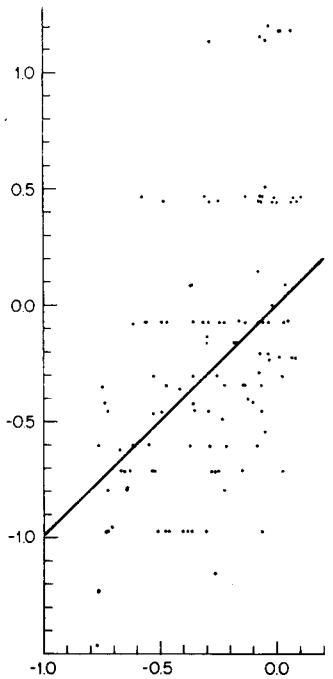


Figure B2

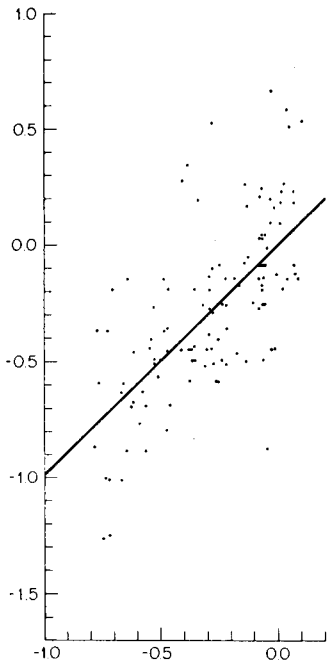


Figure B3

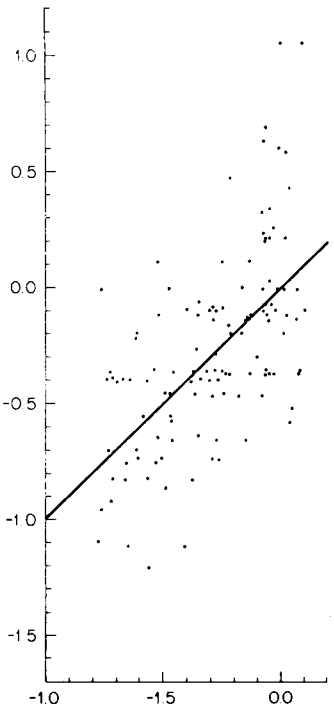


Figure B4

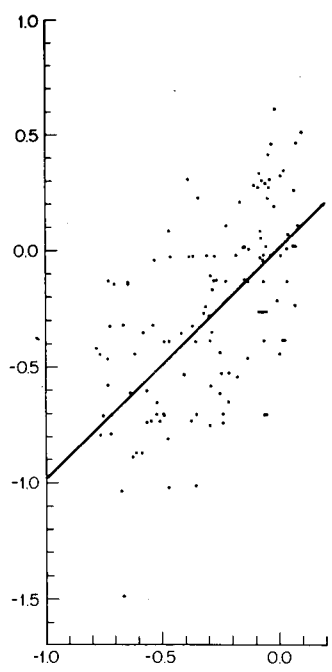


Figure B5

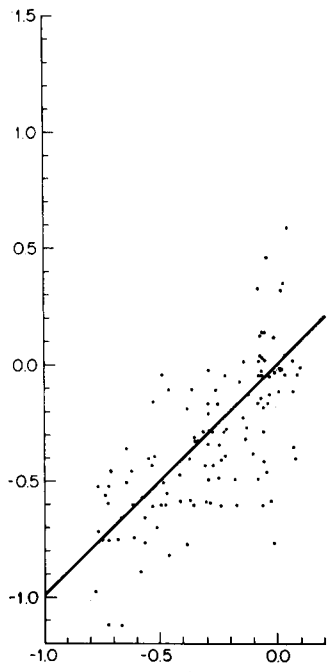


Figure B6

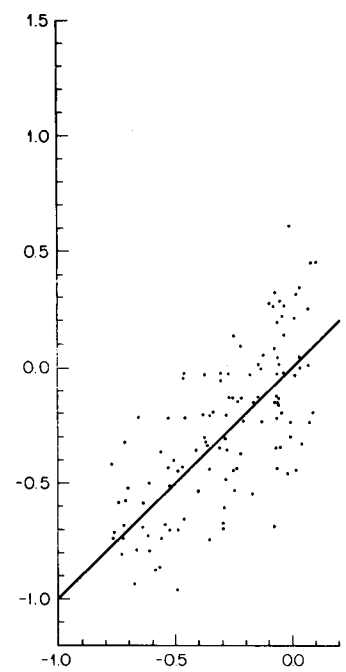


Figure B7