Effects of Individual Optimization in Setting the Boundaries of Dichotomous Items on Accuracy of Estimation

Fumiko Samejima
University of Tennessee

Applying the normal ogive model of latent trait theory, two sets of data, simulated and empirical, were analyzed. The objective was to determine how much accuracy of estimation of the subjects' latent ability can be maintained by tailoring for each testee the order of presentation of the items and the border of dichotomization for each item. This was compared to the information provided by the original graded test items. Results indicated that tailored testing is promising especially when the number of items is not too small, and that a graded item can effectively be used as the initial item in tailored testing because of its branching effect.

The rationale and contents of study follow latent trait theory with a unidimensional latent space and the graded response level (Samejima, 1972). Let Θ be the uni-dimensional latent trait, or latent variable, such that

$$-\infty < \theta < \infty \ . \qquad [1]$$

Let x_g be the item response category or item score for item g, which assumes non-negative integers 0 through m_g. Let $P_{x_g}(\Theta)$ be the operating characteristic of item score x_g, or the conditional probability that the examinee gets item score x_g, given Θ. The response pattern of an individual examinee is given by a vector of order n, the number of items in the test, such that

$$V' = (x_1, x_2, \ldots, x_g, \ldots, x_n) \ . \qquad [2]$$

It is assumed that the distributions of item scores x_g for $g = 1, 2, \ldots, n$ are independent for any fixed value of Θ, so that the operating characteristic of the response pattern V is given by

$$P_V(\theta) = \prod_{x_g \in V} P_{x_g}(\theta) \ . \qquad [3]$$

It should be noted that this operating characteristic is also the likelihood function for estimating the individual parameter Θ, given the response pattern V. The item response information function $I_{x_g}(\Theta)$ is given by

$$I_{x_g}(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_{x_g}(\theta) \ , \qquad [4]$$

and the item information function is the conditional expectation of the item response information function, given Θ, such that

$$I_g(\theta) = E[I_{x_g}(\theta)] = \sum_{x_g=0}^{m_g} I_{x_g}(\theta) P_{x_g}(\theta) \ . \qquad [5]$$

The response pattern information function $I_V(\Theta)$

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 1 Winter 1977 pp. 77-94

is given by

$$I_V(\theta) = \sum_{x_g \epsilon V} I_{x_g}(\theta) \qquad [6]$$

and the test information function is defined equivalently by

$$I(\theta) = \sum_{V} I_V(\theta)\, P_V(\theta) = \sum_{g=1}^{n} I_g(\theta) \qquad [7]$$

(Samejima, 1969, Chapter 6). It has been shown by Samejima that if we add another category boundary to make the number of item response categories $m_g + 2$, and so on, under a general condition the item information function assumes a higher value at any point of Θ.

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 (1957), and the proof has been shown in a simplified form by Kendall and Stuart (1961, Chapter 18). More importantly, it has been shown that in such a case, the asymptotic distribution of the maximum likelihood estimate $\hat{\Theta}$ is normal, with the true value Θ as the mean and

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

as the variance. In latent trait theory, this corresponds to the situation where all the items have the same number of response categories and the corresponding sets of operating characteristics are identical, and the variance of the asymptotic normal distribution is $1/I(\Theta)$. This situation is not likely to occur in practice, however, and in most cases we do not use such a set of equivalent items. The proof for this theorem can be expanded easily under a general condition to the situation where the items are not equivalent (Samejima, 1975). Thus in our situation the test information function has an important role in maximum likelihood estimation, i.e., it serves as a measure of accuracy in the estimation of Θ.

The idea of individualized adaptive or tailored testing (see Weiss & Betz, 1973 for a review) has been developed from the theoretical observation that a test with fewer items can estimate each testee's ability level more accurately than can an ordinary uniform test. This test would contain an optimal set of items, selected from a larger set of test items, whose accuracies of estimation of Θ are greatest at the testee's ability level. Although classical test theory has been used in adaptive testing research, such research will be most effective if based on modern test theory or latent trait theory. In fact, the idea of "tailoring" an optimal set of test items for an individual testee has appeared clearly in some earlier theoretical work in latent trait theory (e.g., Samejima, 1969, Chapter 9). Lord has published many papers (e.g., Lord, 1971, 1972) in which he emphasized the usefulness of latent trait theory and item characteristic functions.

When the model satisfies the unique maximum condition (Samejima, 1969, 1972), the estimation of Θ from the subject's response pattern can be conducted with the aid of a relatively simple computer program, either on the graded or dichotomous response level. This is done by having the computer search the value of Θ which satisfies

$$\sum_{x_g \epsilon V} A_{x_g}(\theta) = 0 \,, \qquad [8]$$

where the basic function $A_{x_g}\Theta$ is defined by

$$A_{x_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{x_g}(\theta) \,. \qquad [9]$$

In individualized adaptive testing, this process will be repeated after presentation of each new item, until the value of the maximum likelihood estimate is reasonably stabilized. To provide rapid convergence of the maximum likelihood estimate, the selection of optimal items will be by far the most important, and the item information function can effectively serve this purpose. Let the computer present the initial item, which has substantial values of information for a wide range of Θ, by which the subject's true

ability level is likely to be covered. If the first estimate of Θ for the subject turned out to be a finite value, then let the computer present a second item, selected from the remaining items, having the maximum amount of information at that level of Θ. If the first estimate of Θ turned out to be either positive or negative infinity, then let the computer present a second item having the maximum amount of information for the interval of higher or lower values of Θ, depending upon the situation, and repeat the process until the estimate of Θ assumes a finite value. From then on, an optimal item is always the one having the maximum amount of information at the level of Θ that the present maximum likelihood estimate assumes. This process is repeated until the value of the maximum likelihood estimate becomes reasonably stabilized.

As mentioned before, tailored testing can be conducted either on the dichotomous or graded response level of test items. We can say, however, that it will be most effective when only a set of dichotomous test items is available, because in such a case the item information function for each item tends to be great only for a very limited interval of Θ (Birnbaum, 1968, Chapter 20; Samejima, 1969, Chapter 6). In such a case, a uniform test tends to have intervals at which the amount of test information is small, and therefore the accuracy of estimation of Θ on these levels of Θ is poor, i.e., the phenomenon known as the attenuation paradox (e.g., Lord and Novick, 1968). If each item has three, four, or more score categories, the amount of information given by the item tends to be great for a wide range of Θ, and therefore the attenuation paradox is less likely to occur.

We must note, however, that even in such a situation where a large set of graded test items is available, individualized adaptive testing will provide us with more accurate estimation of Θ. This probably can be done with fewer test items for each subject, compared to an ordinary test with a uniform set of items. The decision whether to choose tailored testing or conventional testing in such a situation will depend on such factors as one's budget and the availability of well-trained instructors.

Purpose

The purpose of the present study was to find out how much accuracy of estimation can be maintained if each item is dichotomized in such a way that the selection of items and borders of dichotomization are adapted for each subject. This would be compared to the estimation using the full information given by the graded test items. Suppose that we have n items, each with more than two score categories, $m_g + 1$. We always can make a dichotomous item from each graded item by rescoring it using one of the m_g category boundaries for dichotomization. In so doing, we necessarily will lose some amount of information. But we can minimize this loss by selecting the best boundary of the m_g, best in the sense that the dichotomization gives the greatest possible accuracy around the examinee's ability level. This should be done for each subject, because there is no way to set a single set of boundaries for the group of subjects unless their ability levels are very close to one another.

Method

There are two sets of data, one simulated and the other empirical. There are a certain number of hypothetical or actual subjects, each having a response pattern of n graded item scores. The maximum likelihood estimate for each subject is obtained from this response pattern, using the normal ogive model such that

$$P_{x_g}(\theta) = \int_{a_g(\theta - b_{x_g+1})}^{a_g(\theta - b_{x_g})} \psi(u)\, du , \quad [10]$$

where

$$\psi(u) = [2\pi]^{-1/2} \exp[-u^2/2] \quad [11]$$

(Samejima, 1969). Then the same data are reanalyzed as if the test were given in the individualized adaptive testing situation using di-

chotomous items. A program was written for this purpose, in which the computer presents an initial item selected by the experimenter. From then on it selects an optimal item and the boundary of dichotomization in such a way that the information given by the dichotomized item is maximal at the estimated Θ, which was obtained by maximum likelihood estimation in the previous step. The process is repeated until all the n items are used with such dichotomizations, and the resulting maximum likelihood estimate for each subject is observed in comparison with the one obtained earlier using full information of graded items. For convenience, hereafter we shall call this second procedure the *simulated tailored testing situation.*

Results on Simulated Data

In this part of the study, 24 hypothetical test items with $m_g = 3$ for all g's are used. Table 1 shows the item parameters a_g and b_{x_g} for $x_g = 1$, 2, 3 for these items. These parameter values are intended to be realistic, considering the configuration of those for the actual NMB subtest, which will be introduced in the next section. The test information function was computed using

TABLE 1
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	0.5	-0.7	-0.5	0.2
2	0.5	-2.0	-0.8	-0.2
3	0.6	0.3	0.8	2.1
4	0.6	0.0	0.4	1.3
5	0.7	-1.3	-0.2	0.4
6	0.7	0.2	0.9	2.0
7	0.8	-0.5	0.8	1.9
8	0.8	-1.1	-0.9	-0.1
9	0.9	-0.2	0.4	0.6
10	0.9	-1.6	-1.0	0.2
11	1.0	-1.8	-1.1	-0.6
12	1.0	0.1	1.4	1.6
13	1.1	-0.1	0.8	1.1
14	1.1	-1.0	-0.5	0.0
15	1.2	-1.2	-0.2	0.8
16	1.2	-1.7	-0.8	-0.5
17	1.3	-0.3	0.5	1.4
18	1.3	-0.6	0.4	0.8
19	1.4	-0.9	0.3	1.1
20	1.4	-0.4	-0.1	0.6
21	1.5	-1.9	-1.6	-1.2
22	1.5	-1.5	-0.4	0.9
23	1.6	-0.8	-0.4	0.8
24	1.6	-1.4	-0.6	0.4

Equation 7 in the normal ogive model, and the result is shown in Table 2. Thus it is obvious that this hypothetical test will be most informative, or the estimation of Θ will be most accurate, around $\Theta = -0.3$. For this reason, a set of 100 response patterns was calibrated by the monte carlo method fixing the value of Θ at -0.3. The maximum likelihood estimate of Θ for each response pattern was obtained by a computer program as the value of Θ which satisfies Equation 8.

TABLE 2
Test information function of the hypothetical test of twenty-four graded 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 1 shows the cumulative frequency distribution of these 100 maximum likelihood estimates, $\hat{\Theta}$, along with the dashed curve representing the normal distribution function, $N(\Theta, 1/I(\Theta))$. The standard deviation of this distribution was approximately .2128. From this result, it is clear that with the number of items as small as 24 and the value of m_g for each item as small as 3, the cumulative frequency distribution of $\hat{\Theta}$ is already very close to the normal distribution function, which is the asymptotic distribution of $\hat{\Theta}$ when n tends to positive infinity, although the example is only of one sample of size 100.

As a next step, each of the 24 items was rescored dichotomously in two different ways, one giving the greatest and one the least possible information at $\Theta = -0.3$. Figure 2 presents the two cumulative frequency distributions. The dashed curves represent the normal distribution functions, whose standard deviations are the square roots of the reciprocals of $I(\Theta)$, which were .2407 and .3685 respectively. It will be observed that the fits of the curves to the cumulative frequency ratios are not as good as the one for the graded items, and yet they are fairly good. The category borders used for the rescoring are, for items 1 through 24, 2, 3, 1, 1, 2, 1, 1, 3, 1, 3, 3, 1, 1, 2, 2, 3, 1, 1, 1, 1, 3, 2, 2, and 2 for the most informative dichotomization, and 3, 1, 3, 3, 1, 3, 3, 1, 3, 1, 1, 3, 3, 1, 3, 1, 3, 3, 3, 3, 1, 3, 3, and 1 for the least informative dichotomization. Since the first set of dichotomized items was "tailored" for the ability level $\Theta = -0.3$, the discrepancy between .2407 and .2128 in the standard deviation can be considered as the difference in accuracy of estimation between the set of graded items and the set of best possible dichotomized items.

A question may arise: What will be the accuracy of estimation if we tailor a set of dichotomized items specifically for each hypothetical subject, instead of giving a uniform set of dichotomized items? To answer this question, maximum likelihood estimation was made in the simulated tailored testing situation described above, using nine different initial items, of which seven were dichotomous and two were graded. Table 3 presents these initial items, followed by the category borders used for dichotomization for the first seven items, and the values of their item information functions at $\Theta = -0.3$. In each case, the test information function was different for different response patterns, since a

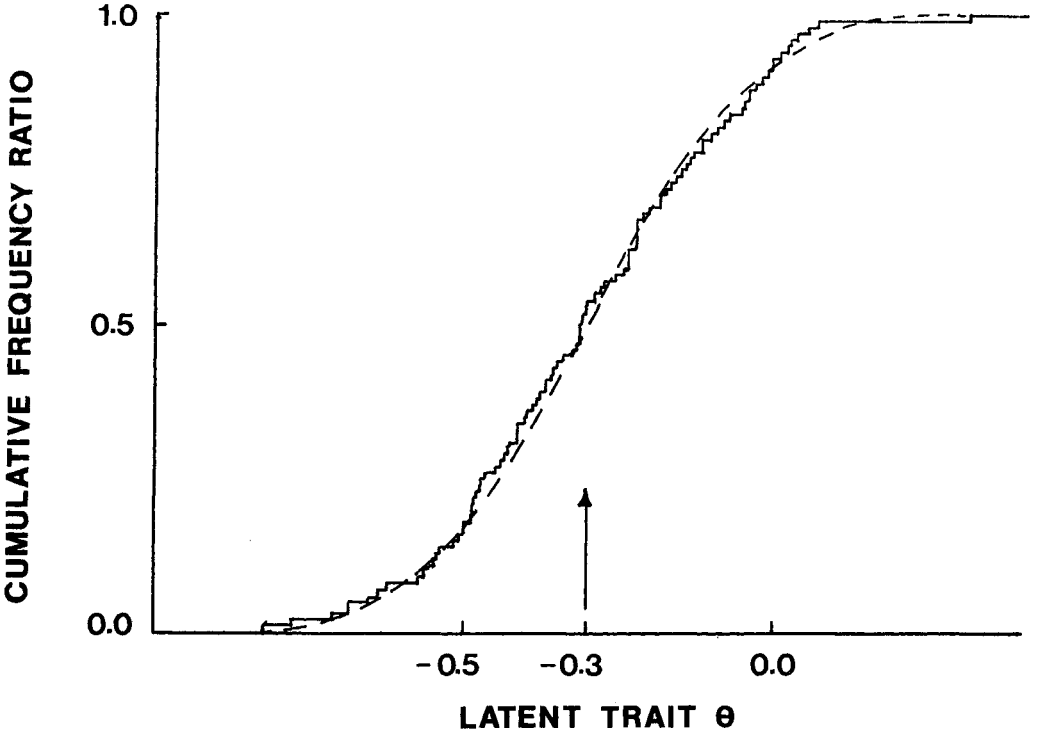


FIGURE 1

Cumulative frequency ratio of maximum likelihood estimates based on the response patterns of graded item scores for the one hundred hypothetical subjects with the ability level –0.3 (solid line) and the normal distribution function, $N(-0.3, 0.2128^2)$, (dashed curve).

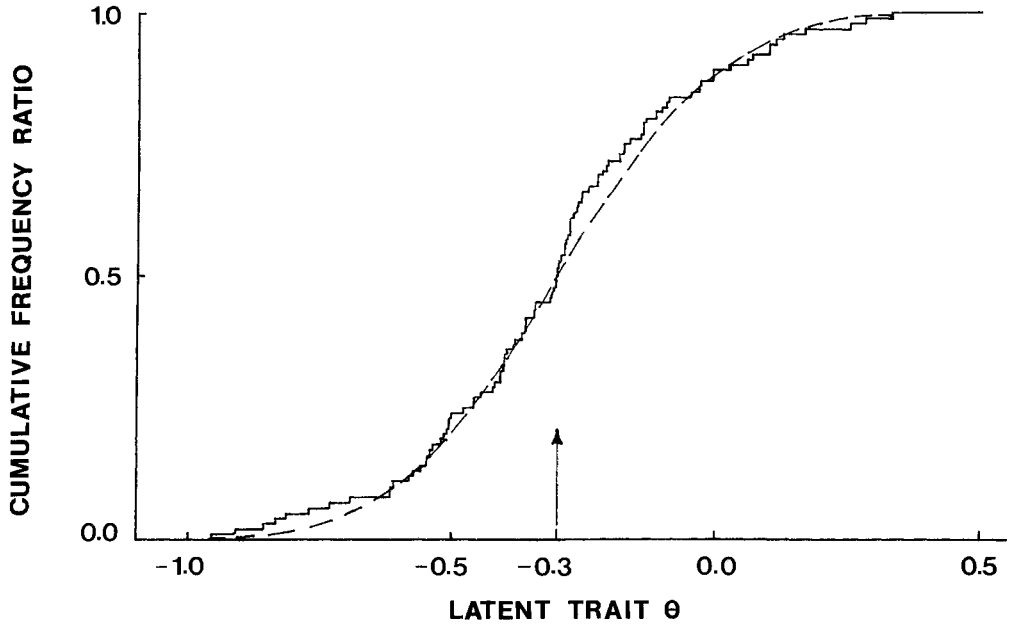


FIGURE 2

Cumulative frequency ratios of maximum likelihood estimates based on the response patterns of dichotomous scores, using the most informative dichotomization (above) and the least informative dichotomization (below) at the ability level –0.3, for the same one hundred hypothetical subjects (solid lines) and the normal distribution functions (dashed curves), $N(-0.3, 0.2407^2)$ (above) and $N(-0.3, 0.3685^2)$ (below).

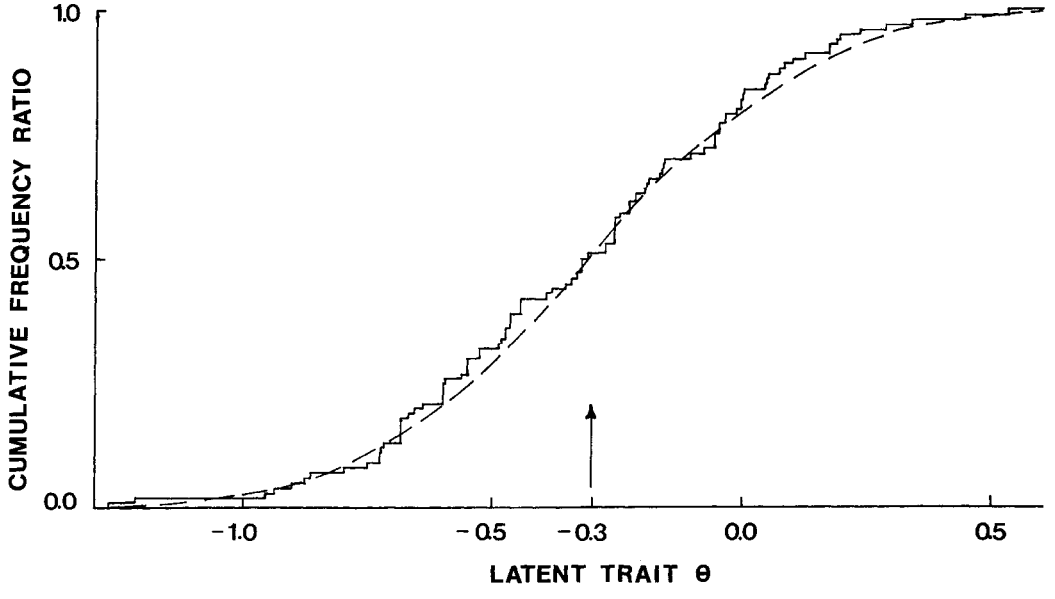


TABLE 3

Mean square errors and other indices for the variability of the maximum likelihood estimates in the simulated tailored testing situations, in which nine different initial items are used.

	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

set of dichotomized items was tailored for each subject. Therefore, there is no single index comparable to the standard deviation of the normal distribution obtained in each of the previous three cases. For this reason, the mean square error, i.e., the mean of the squared deviation of each maximum likelihood estimate from −0.3, was computed, and its square root was used as the index comparable to the standard deviations in the previous three cases. It will be observed that in all cases this index shows a value reasonably close to .2407 of Case 2; some are even less, and they all are far less than .3685 of Case 3. In actual situations, there is no way of knowing the true value of Θ for any subject, and therefore there is no way of tailoring a set of items for the subject's true ability level. The present result therefore shows good promise for individualized adaptive testing, which is based on information given by the subject's previous responses to the items.

Figure 3 shows how the cumulative frequency distribution of the maximum likelihood estimates, obtained in the simulated tailored testing situation, deviates from $N(-0.3, 0.2128^2)$, for two examples in which the values of the item information of the initial items 3-3 and 14-3 are substantially different, and so are the values of the mean square error (cf. Table 3). In the first case discrepancies between the two curves are substantially large, while the fit is good for the second example.

In the above result, it turned out that the best set of maximum likelihood estimates in the simulated tailored testing situation was provided in the case where the dichotomized item, 14−3, with a moderate value of item information function at $\Theta = -0.3$, was used as the initial

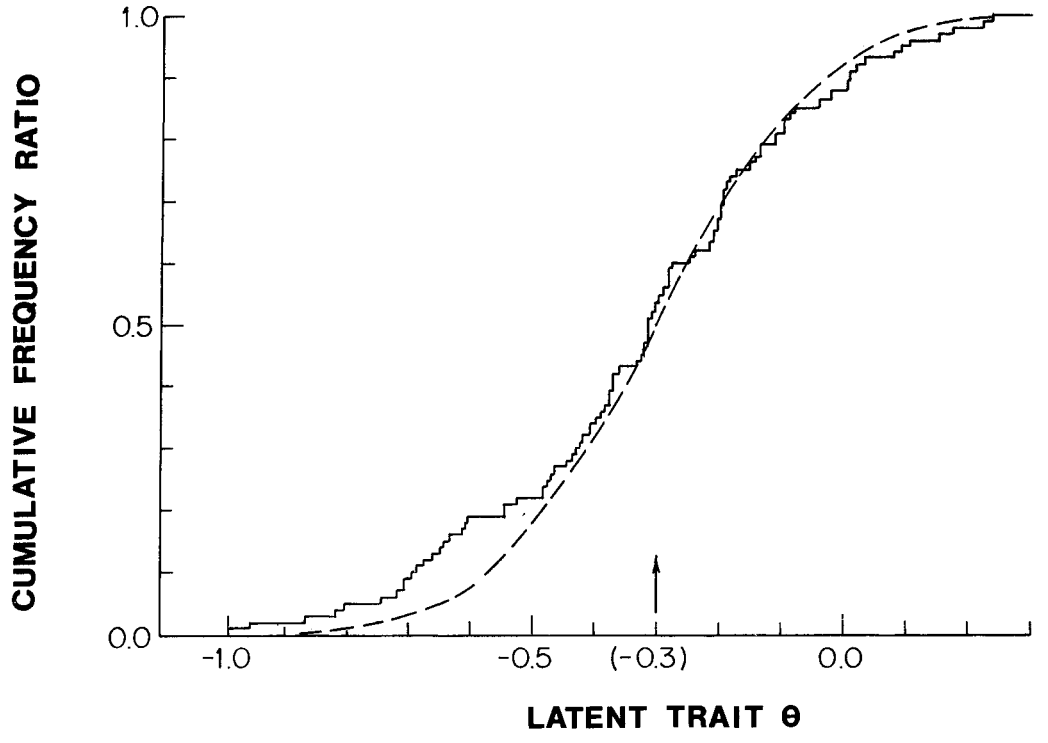
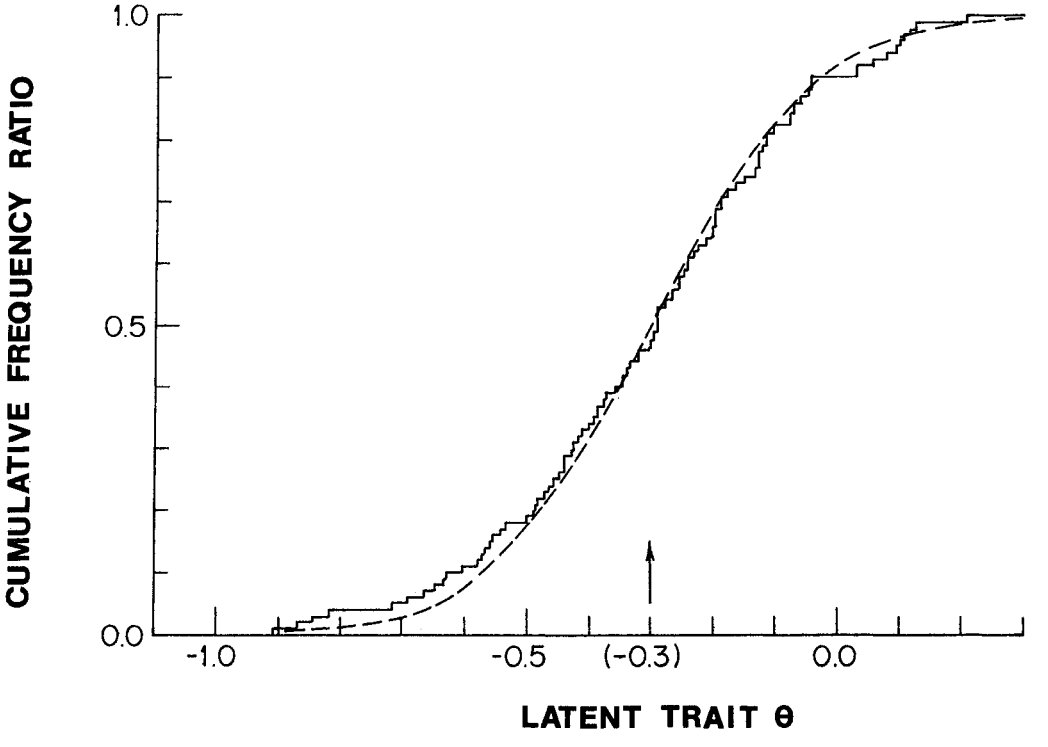


FIGURE 3

Cumulative frequency ratios of maximum likelihood estimates obtained by simulated tailored testing with two different initial items, 3–3 and 14–3, respectively, for the same one hundred hypothetical subjects and the normal distribution function, N(–0.3, 02128²).



item. The next best was the case in which the most informative graded item, 24, was used as the initial item. To make a more precise comparison of these two cases, the mean square error was computed for the set of maximum likelihood estimates after each of the 4th, 6th, 8th, 12th, 16th, 20th and 24th presentations of items in each case. Figure 4 presents the square roots of the mean square errors thus obtained. It will be observed that there are substantial differences between these two values at the earlier stages of simulated tailored testing. We can consider this as an indication of the branching effect provided by a graded item, and further investigations in this direction in tailored testing will be valuable. This branching effect seems to disappear in the later stages. Note, however, that this simulated tailored testing situation is different from actual individualized adaptive testing situation in two ways: 1) all the 100 response patterns were calibrated at $\Theta = -0.3$;2) only the ways of dichotomization of the items and their order of presentation were tailored, using the whole set of 24 items, while in the actual tailored testing situation the items are selected from the whole set containing a much larger number of items. Thus it is likely that especially this second factor negatively affects the branching effect of the initial graded item.

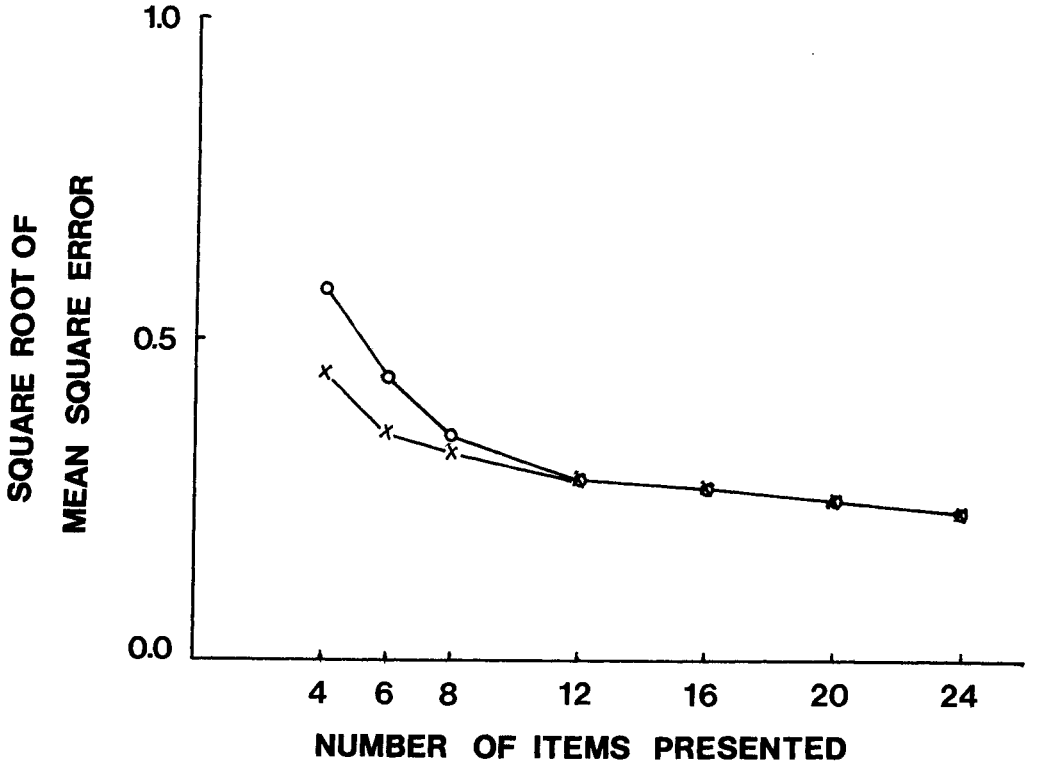


FIGURE 4

Comparison of the two sets of the square roots of the mean square errors of the maximum likelihood estimates obtained by simulated tailored testing with two different initial items, 24 (x) and 14–3 (o), computed after presenting 4, 6, 8, 12, 16, 20 and 24 items respectively.

Results on Empirical Data

In this section, observations were made on empirical test data collected for 446 subjects, mostly college and summer school students in the United States and Canada in March through July, 1974. The test used in this project was developed for research purposes and consists of 18 items, of which ten belong to the figural subtest (FGR) and eight to the numerical subtest (NMB) (Samejima, 1975). One of the ten FGR items was omitted from the analysis, for it turned out to be too difficult and most subjects could not solve it. Of the 446 subjects, 40 were excluded from the analysis because they failed in completing one or more items. The test was administered by a well-trained instructor to the remaining 406 subjects in groups of up to 36 subjects, although in many cases there were fewer than 10. In some sessions FGR was presented first, and in others NMB was presented first. Each session required approximately 90 minutes, including the initial instructions and a five-minute break between the two subtests. A time limit of two to six minutes was set for each item, with the exception of the eighth item of NMB, N-8 whose time limit was thirteen minutes. The instructor called time twice for each item, once when there was one minute left and once when the time was up. Each item was scored in a graded way with $m_g = 3$ for all the items of FGR and the first seven items of NMB, and with $m_g = 7$ for N-8. It turned out, however, that frequencies for some item score categories were too small, and appropriate recategorizations were necessary for some items. For this reason, the number of category boundaries, m_g, was reduced for some items (see Tables 4 and 5), and these recategorizations made the frequency of each score category at least as large as 18.

Multivariate normality was assumed for the 17 item variables, which are assumed behind the 17 sets of item scores, and the polychoric correlation coefficient (Tallis, 1962) was computed for each pair of the item variables using Lieberman's (1969) program. Principal factor analysis was applied to the resulting correlation matrix, and several different rotations uniformly indicated two clusters, one for each subtest (Samejima, 1975). For this reason, principal factor analysis was applied again for each subtest, using the SPSS program with iteratively estimated communalities. The first principal factor for FGR was named Figural Ability, and that for NMB was named Numerical Ability. The eigenvalue for the former turned out to be 3.029 or 60.2% of the sum of the communalities, and that for the latter turned out to be 4.132 or 79.5% of the sum of the communalities.

With the above assumptions, the normal ogive model on the graded response level (Samejima, 1969 and 1972) in the unidimensional latent space can be applied, with the parameter values given by

$$a_g = \rho_g / (1 - \rho_g^2)^{1/2} \qquad [12]$$

and

$$b_{x_g} = \gamma_{x_g} / \rho_g \qquad [13]$$

$$\text{for} \quad x_g = 1, 2, \ldots, m_g \ ,$$

where ϱ_g is the factor loading of item g on the figural or numerical ability and γ_{x_g} is the normal deviate corresponding to the proportion of the subjects who received the item score x_g or greater. These parameter values are presented as Tables 4 and 5 respectively.

Since there is no way to know the true ability level of the individual subject with empirical data, it is impossible to make the kind of observation given in the preceding section. One alternative may be to use the maximum likelihood estimate, $\hat{\Theta}$, which is obtained on the full information given by the graded items, as a substitute for the true ability score. This type of observation was made for the 123 subjects for FGR and for the 138 subjects for NMB elsewhere (Samejima, 1975). The result indicates a more accurate estimation of Θ in simulated tailored testing with various initial items in comparison to the case in which each item is uniformly dichotomized. Also indicated is the branching effect of the graded item when it is

TABLE 4
Item parameters of the nine test items of the subset FGR.

Item	Discrimination Index	Difficulty Indices b_{x_g}		
g	a_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 5
Item parameters of the eight test items of the subtest NMB.

Item	Discrimination Index	Difficulty Indices b_{x_g}			
g	a_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

used as the initial item in simulated tailored testing.

Another interesting observation may be the comparison of the frequency distributions of the maximum likelihood estimates for the 406 subjects, obtained by different methods. The theoretical density function of the maximum likelihood estimates, $g(\hat{\Theta})$, will be given by

$$g(\hat{\theta}) = \int_{-\infty}^{\infty} \phi(\hat{\theta}|\theta)\, f(\theta)\, d\theta \quad , \quad [14]$$

where $f(\Theta)$ is the density function of Θ and $\phi(\hat{\Theta}|\Theta)$ is the conditional density of $\hat{\Theta}$ given Θ. Under the present conditions, $f(\Theta)$ is the standard normal density function and $\phi(\hat{\Theta}|\Theta)$ can be approximated by the density function of $N(\Theta, 1/I(\Theta))$, although the number of items in each subtest is fairly small. Figure 5 presents the histogram representing the theoretical frequency distribution of $\hat{\Theta}$ obtained from $g(\hat{\Theta})$, using the test information function of the original graded

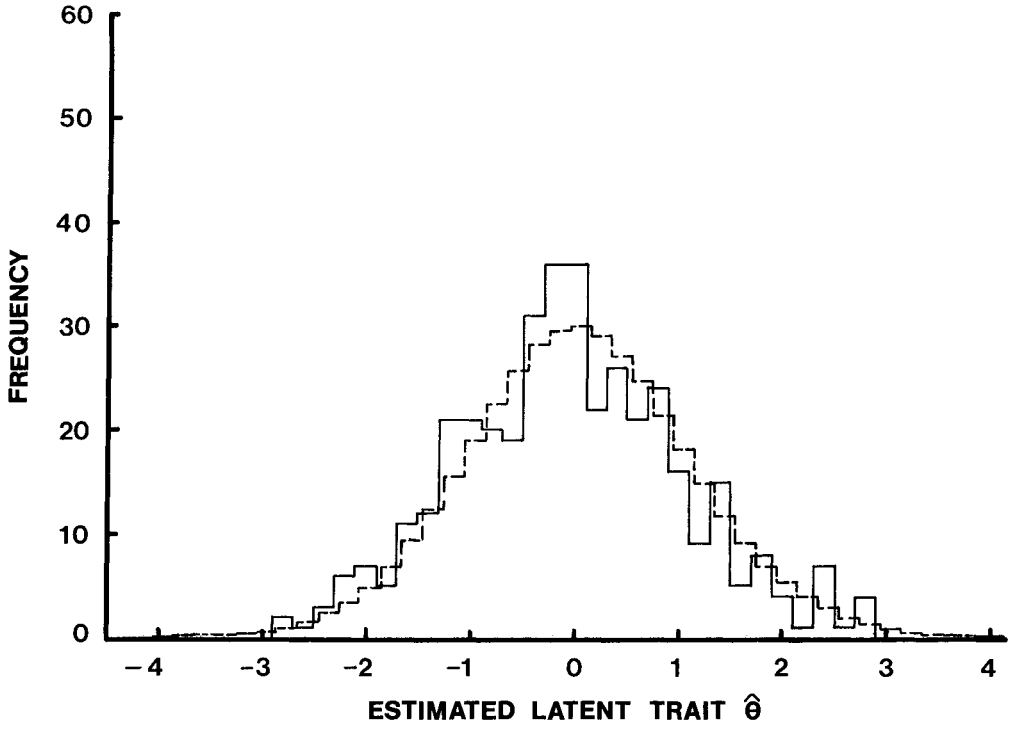
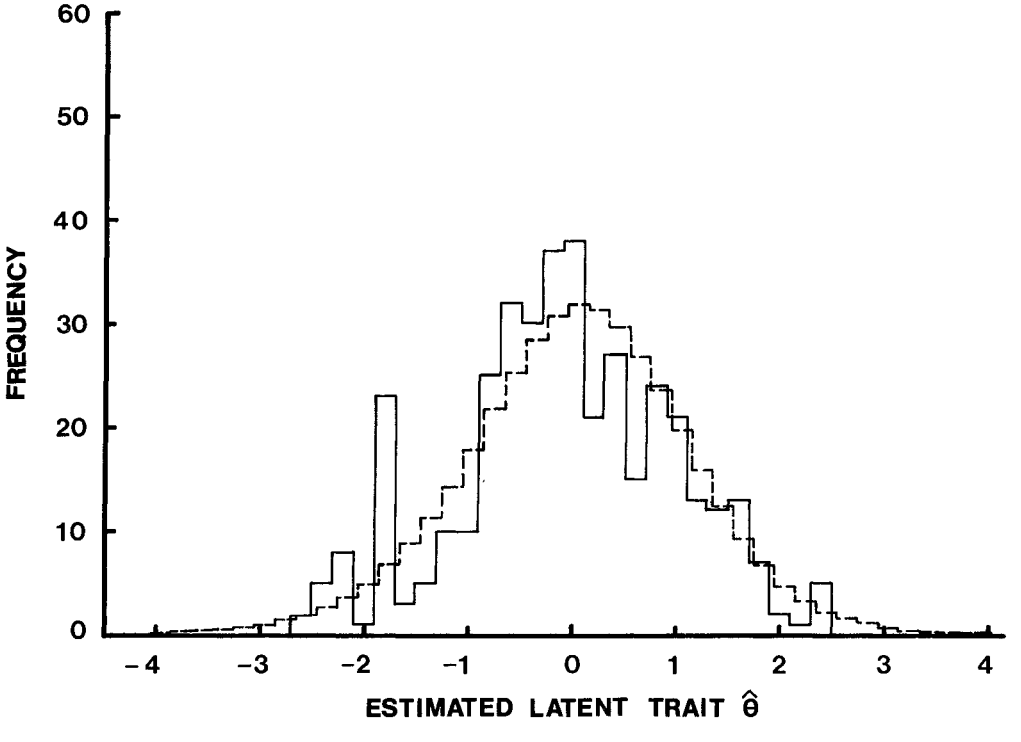


FIGURE 5

Comparison of the histograms representing the theoretical frequency distribution of the maximum likelihood estimate of Θ (solid line) obtained from the test information function of the original graded items and its actual frequency distribution (dashed line), for the subtest FGR (above) and NMB (below).



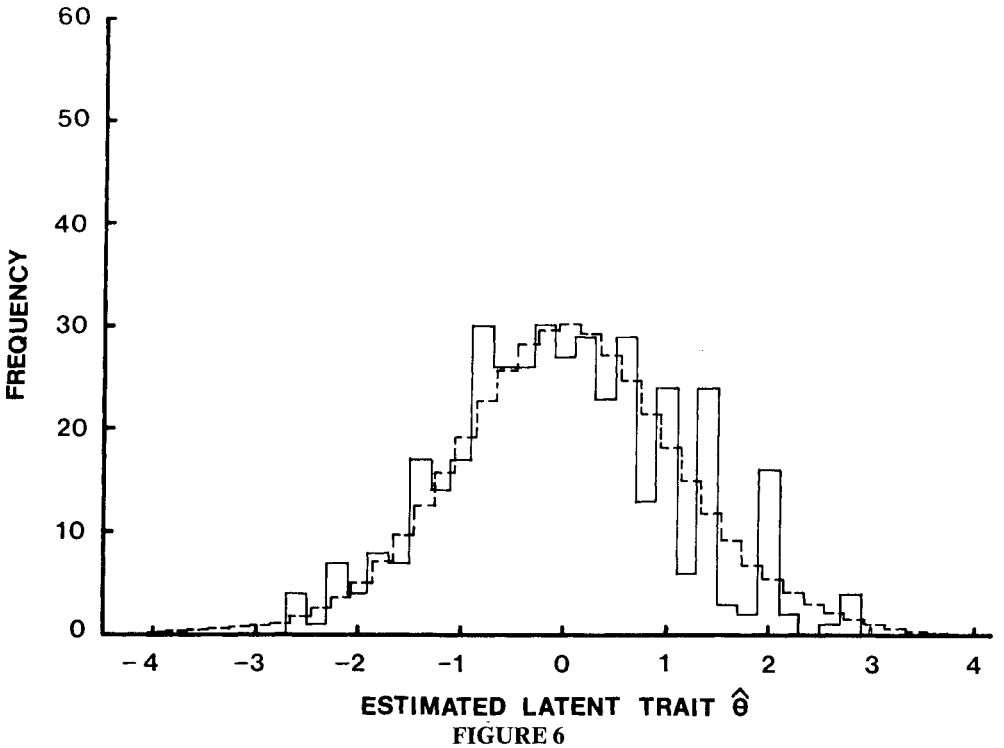
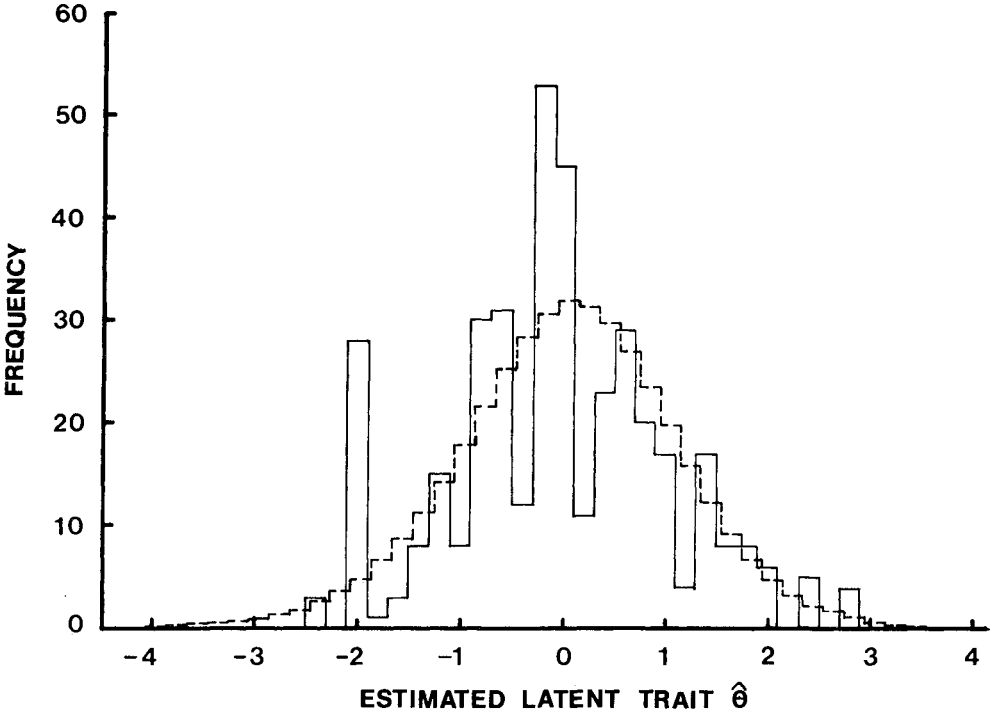


FIGURE 6

Comparison of the histograms representing the same theoretical frequency distribution used in Figure 5 and the actual frequency distribution of the maximum likelihood estimate (dashed line) obtained in the simulated tailored testing situation, for two extreme cases in which the initial items are F–3 and N–1–2 for the FGR and NMB subtests respectively.



items in Equation 14, and the actual frequency distribution of $\hat{\Theta}$ (dashed histogram) for each of the subtests FGR and NMB. The value of the test statistic for the goodness of fit of the frequencies was computed to provide 17.458 for FGR and 34.388 for NMB, with an appropriate grouping of some categories to make each theoretical frequency no less than 8, which left 21 categories for each subtest. The number of degrees of freedom is unknown. If we take 20, however, $\chi^2_{0.02} = 35.020$ and $\chi^2_{0.05} = 31.410$, although the true value will be less than 20. It may be said from these results that the fit is fair in both cases, and it is especially so for FGR.

Because the test information function is greatest for all Θ when graded scoring is used (Samejima, 1969, Chapter 6), and because the fit of the theoretical frequency distribution of $\hat{\Theta}$ to the actual frequency distribution was fair for both FGR and NMB, it will be legitimate to use the same theoretical frequency distribution to find out the accuracy of estimation of Θ in simulated tailored testing, for each of the subtests FGR and NMB. Figure 6 presents two examples of ten such comparisons of the theoretical and actual frequency distributions, in which the actual frequencies are based on the maximum likelihood estimates obtained in the simulated tailored testing situation, using as the initial items F-3 for the FGR Subtest and N-1-2 for the NMB Subtest. The initial items used for FGR are F-2-2, F-6-2, F-3-3, F-2 and F-3 respectively, and those used for NMB are N-7-2, N-3-1, N-1-2, N-7 and N-4 respectively. The second number in these designations, if it exists, indicates the border for dichotomization for the first three items. The values of the test statistic were 120.749, 91.462, 46.848, 46.415 and 40.957 for FGR, and 72.649, 119.382, 125.249, 75.270 and 68.192 for NMB. The two examples show the best and worst fits, i.e., the cases in which F-3 and N-1-2 are used as the initial items respectively.

From this result, it is obvious that all these values are substantially larger than the corresponding values of the test statistic when the actual frequencies of the maximum likelihood estimates were obtained from the response of graded items scores, i.e., 17.458 for FGR and 34.388 for NMB. This result indicates relatively poor accuracies of estimation of Θ in the simulated tailored testing situation. This overall tendency may be caused by the fact that in each subtest the number of items is small, much smaller than the 24 used in the study of simulated tailored testing data in which the accuracy of estimation of Θ proved to be fair. Note, however, that the values of the test statistic in the last two cases of each subtest, where a graded item was used as the initial item, are less than most of the other three cases where a dichotomous item was used. These two graded items are the most informative and the second most informative items for the interval of Θ, [-0.8, 0.1] for FGR, and for the interval of Θ, [-0.1, 1.0] for NMB. The above result may be interpreted as the branching effect of the graded item in preference to the dichotomous item.

Similar comparisons of the actual frequencies of the maximum likelihood estimates with the theoretical frequencies were made for two additional cases, in which the estimates were obtained from the response patterns of uniformly dichotomized item scores for each subtest. The selection of the boundaries of dichotomization was made in such a way that the resulting test information function is the greatest for the interval of Θ, [-0.8, 0.1] in the first case, and it is the least for the same interval in the second case, for FGR. The same principle was used for NMB using the interval of Θ, [-0.1, 1.0]. Figure 7 presents these test information functions, along with the test information function of the original graded items for each subtest. As we can see in this figure, these intervals of Θ are those in which the original test information functions are greatest. Comparisons were made between the previously used actual frequencies of the maximum likelihood estimates and the theoretical frequencies in these two cases for each subtest. The values of the test statistic were 90.862 and 320.874 respectively for FGR, and 133.507 and 263.217 respectively for NMB. Figure 8 presents the two cases for FGR, in which the contrast is

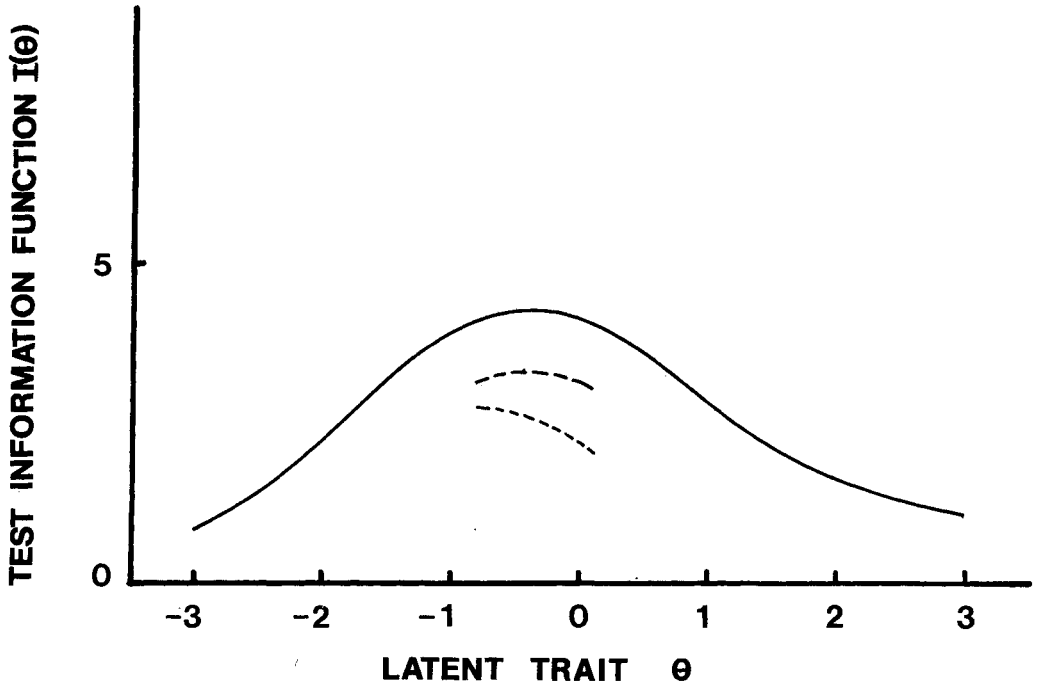
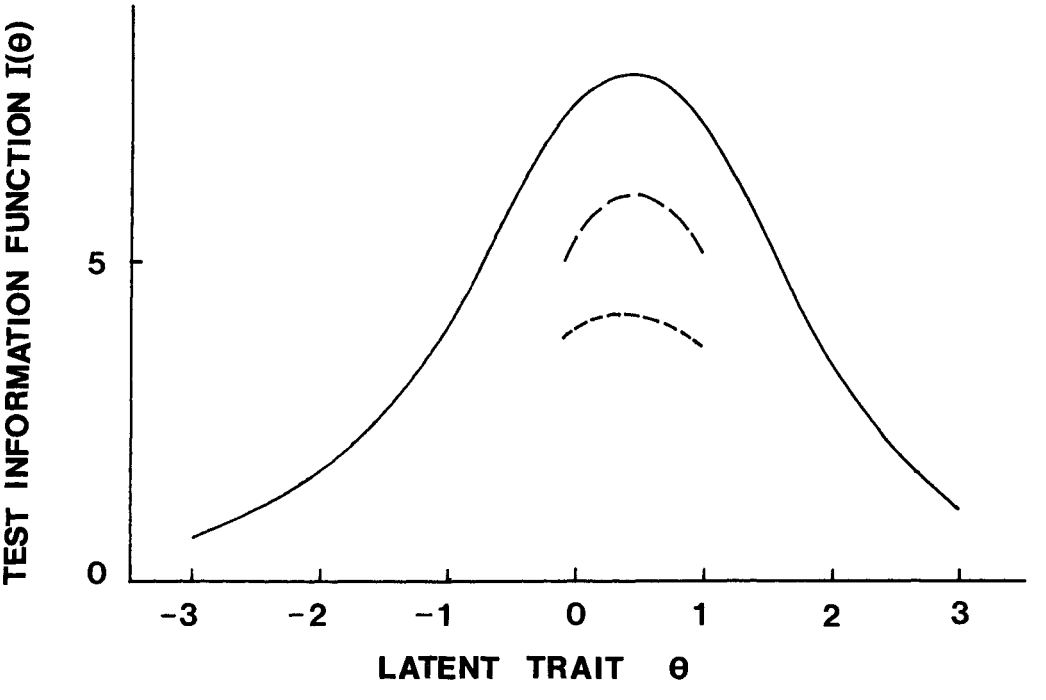


FIGURE 7

Test information functions of the original graded items (solid curve), of the most informative dichotomous items (dashed curve) and of the least informative dichotomous items (dotted curve) for the interval of Θ, [−0.8, 0.1] for FGR (above), and for the interval [−0.1, 1.0] for NMB (below).



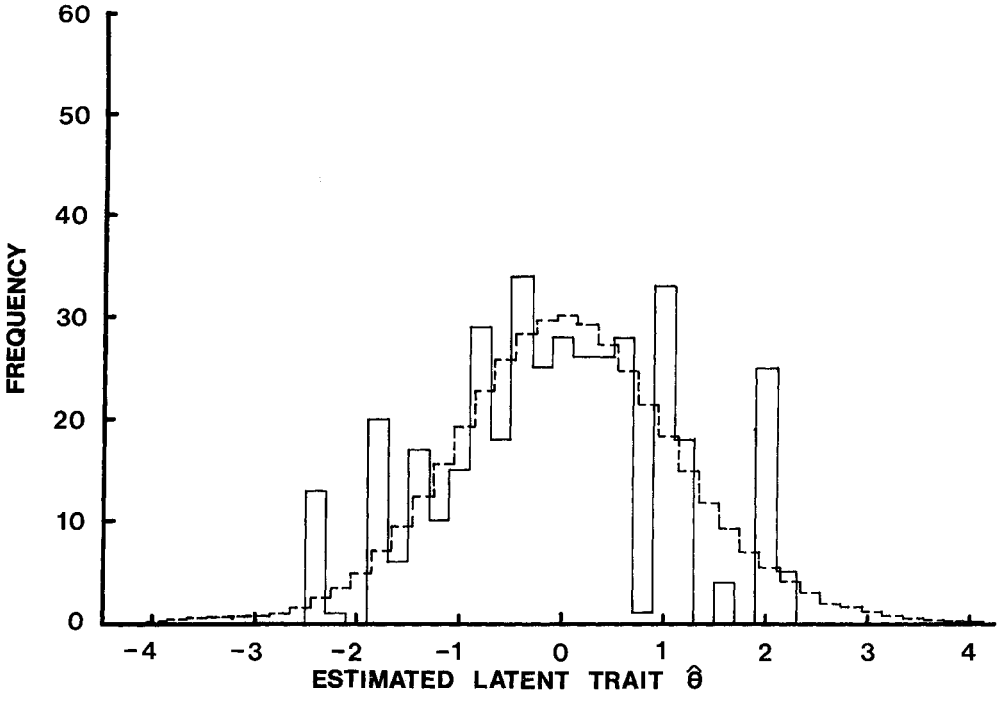
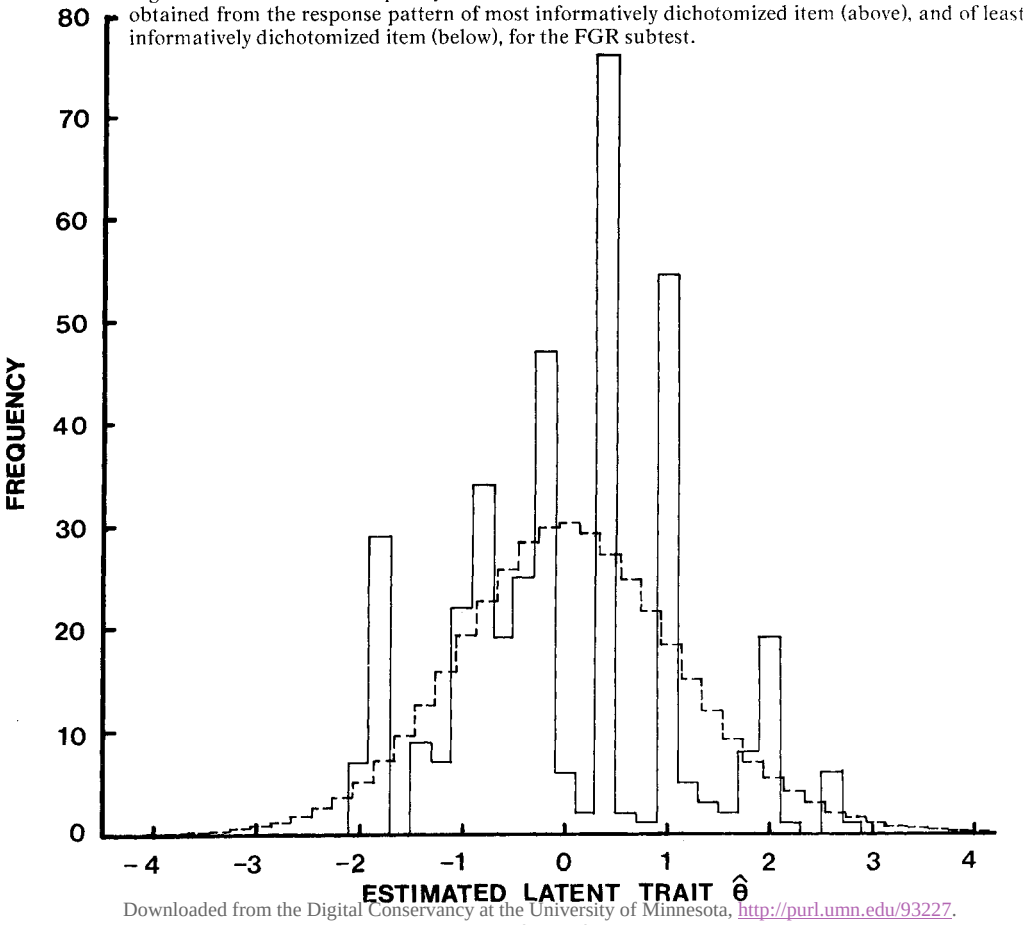


FIGURE 8 Comparison of the histograms representing the same theoretical frequency distribution used in Figure 5 and the actual frequency distribution (dashed line) of the maximum likelihood estimate obtained from the response pattern of most informatively dichotomized item (above), and of least informatively dichotomized item (below), for the FGR subtest.



greater. As was expected, for both subtests, the accuracy of estimation of Θ can be considered substantially better for the case of the most informative dichotomization, compared with the case of the least informative dichotomization. The result also indicates the advantage of simulated tailored testing when an informative graded item is used as the initial item over the case of uniform dichotomization in the accuracy of estimation of Θ.

The three initial dichotomous items used in the first three cases of the simulated tailored testing (i.e., F-2-2, F-6-2 and F-3-3 for FGR and N-7-2, N-3-1 and N-1-2 for NMB) are the most informative, the least informative and a medium informative dichotomous items respectively, for the interval [-0.8, 0.1] for FGR and for [-0.1, 1.0] for NMB. The values of the test statistic, however, do not indicate the effect of the differences in information given by the initial items, as far as these dichotomous items are concerned.

Discussion and Conclusion

The research was designed and conducted for the purpose of determining the effect of individualized adaptive testing on the accuracy of estimation of the latent variable Θ, by reanalyzing the same data in several different ways. In so doing, the simulated tailored testing situation was used for both empirical and simulated data. The result provided some encouragement in using dichotomous items in individualized adaptive testing as well as in using a graded item as the initial item to produce an efficient initial branching effect.

There is no question about the advantage of using graded items over dichotomous items in the accuracy of estimation of Θ in any testing, and therefore it will be more desirable to use graded items throughout individualized adaptive testing. Difficulties may arise, however, in developing a large set of graded test items, each of which does not require so much more time than an average dichotomous item does, in many practical situations. For this reason, it may be more realistic to consider developing only a few graded test items in addition to a large number of dichotomous items, and using the graded items as initial items to produce efficient initial branching effects.

References

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

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.

Lord, F. M. A theoretical study of two-stage testing. *Psychometrika,* 1971, 36, 227-242.

Lord, F. M. Individualized testing and item characteristic curve theory. *ETS Research Bulletin* 72-50, 1972, ETS, Princeton, New Jersey.

Lord, F. M. and M. R. Novick. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

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. Graded response model of the latent trait theory and tailored testing. *Paper presented at the Conference on Computerized Adaptive Testing,* 1975, Civil Service Commissions and Office of Naval Research, Washington, D. C.

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.

Weiss, D. J. and N. E. Betz. Ability measurement: conventional or adaptive? *Psychometric Methods Program Research Report* 73-1, 1973. University of Minnesota, Minneapolis, Minnesota.

Author's Address

Fumiko Samejima, Department of Psychology, University of Tennessee, Knoxville TN 37916.