Personnel
Research and
Development
Center

Professional Series 75-6

Proceedings of the First Conference on Computerized Adaptive Testing

WASHINGTON, D.C., June 12 and 13, 1975

United States
Civil Service Commission
Bureau of Policies and Standards

Professional Series 75-6

PROCEEDINGS OF THE FIRST CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

WASHINGTON, D. C., June 12 and 13, 1975

Sponsored by

The Office of Naval Research
and
The Personnel Research and Development Center

Personnel Research and Development Center
U.S. Civil Service Commission
Washington, D. C.
March 1976

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $2.40

Stock Number 006-000-00940-9

PROCEEDINGS OF THE FIRST CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

FOREWORD

The plan for a conference devoted to the state of research in the field of computerized adaptive testing grew out of a suggestion made in late 1974 by Frederic M. Lord of Educational Testing Service. As one of the principal psychometric architects of the latent trait theory of mental abilities, which underlies the work being done in this field, Dr. Lord observed that it was now time to bring together as many as possible of the people doing this work, for an overview of the state of the art. It was then decided that the appropriate sponsors of such a conference were the Navy, whose Office of Naval Research funds computerized adaptive testing projects in military and educational organizations, and the U.S. Civil Service Commission, where psychologists in the Personnel Research and Development Center have been carrying out research in the area for a number of years. Accordingly, representatives of these two offices met in March, 1975 to take the necessary steps to organize the conference. Members of the organizing committee were: Glenn L. Bryan, Director, Office of Naval Research; Marshall J. Farr, Director, Personnel and Training Research Programs, ONR; Joseph L. Young, Assistant Director, PTRP, ONR; William A. Gorham, Director, Personnel Research and Development Center, U.S. Civil Service Commission; Richard H. McKillip, Chief, Research Section, PRDC; Vern W. Urry, Frank L. Schmidt, and John F. Gugel, Personnel Research Psychologists, PRDC.

The principal objectives of the conference were defined as exchange of information, discussion of theoretical and empirical developments, and coordination of research effort. It was decided that the conference should be invitational, because of its highly technical subject matter, and that invitations would be sent to those persons known to be interested in the subject. Nominations were then made of researchers who should be asked to present papers and to act as discussants. From the list of nominations, the committee selected those nominees it believed would represent the broadest range of effort from theory to practical application and would also represent organizations in the public, private, and military sectors. Dr. Lord and Bert F. Green, Jr. of Johns Hopkins University agreed to serve as discussants.

Edmund F. Fuchs was appointed conference coordinator to implement these decisions, and the conference was held as planned on June 12 and 13, 1975, in Washington, D.C. Sixty-eight persons attended. Fourteen papers were read, and the discussants, who had studied the papers in advance, commented upon them.

Informal discussion during and after the conference and replies to a short questionnaire given to the attendees indicated that the objectives were successfully met. In general, attendees felt that a follow-up conference would be desirable, to pursue further the potential of computers for the measurement of human abilities. Two announcements were made at the conference sessions concerning ways of establishing a continuous exchange of information among researchers.

Cynthia L. Clark
Editor

Proceedings of the First Conference on Computerized Adaptive Testing

Washington D.C., June 12 and 13 1975

**Sponsored by the Office of Naval Research
and the
Personnel Research and Development Center**

CONTENTS

DISCUSSION

U.S. Department of the Navy U.S. Civil Service Commission
Office of Naval Research Personnel Research and Development Center

Program

Conference on Computerized Adaptive Testing

June 12 and 13, 1975

Room 2008, New Executive Office Building, Washington, **D.C.** June 12

Morning Session Chairman: Glenn L. Bryan, Office of Naval Research 9:00 a.m.
Welcome, William A. Gorham, U.S. Civil Service Commission
Preliminary Remarks, Glenn L. Bryan, Office of Naval Research

9:15 The Graded Response Model of Latent Trait Theory and Tailored Testing
Fumiko Samejima, University of Tennessee
9:45 Incomplete Orders and Computerized Testing
Norman Cliff, University of Southern California 10:15(coffee break)
10:30 Adaptive Testing Research at Minnesota: Overview, Recent Results, and Future Directions
David J. Weiss, University of Minnesota
11:00 Adaptive Testing Research at Minnesota: Some Properties of a Bayesian Sequential Adaptive Mental Testing Strategy
James R. McBride, University of Minnesota

June 12, Afternoon Session Chairman: William A. Gorham, U.S. Civil Service Commission

1:00 p.m. An Empirical Investigation of Weiss' Stradaptive Testing Model
Brian K. Waters, U.S. Air Force
1:30 Using Computerized Tests to Add New Dimensions to the Measurement of Abilities Which are Important for On-Job Performance: An Exploratory Study
Charles H. Cory, Navy Personnel Research and Development Center
2:00 A Broad-Range Tailored Test of Verbal Ability
Frederic M. Lord, Educational Testing Service
2:45 Bayesian Tailored Testing and the Influence of Item Bank Characteristics
Carl Jensema, Gallaudet College
3:15 Reflections on Adaptive Testing
Duncan N. Hansen, Memphis State University
3:45 Announcements, Robert J. Gettelfinger, Educational Testing Service

June 13, Morning Session Chairman: William A. Gorham, U.S. Civil Service Commission

9:00 a.m. Computer Assisted Testing: An Orderly Transition From Theory to Practice
Richard H. McKillip, U.S. Civil Service Commission
9:30 Five Years of Research: Is Computer Assisted Testing Feasible?
Vern W. Urry, U.S. Civil Service Commission 10:00 (coffee break)
10:15 Effectiveness of the Ancillary Estimation Procedure
John F. Gugel, U.S. Civil Service Commission
10:45 Item Parameterization Procedures for the Future
Frank L. Schmidt, U.S. Civil Service Commission)

June 13, Afternoon Session Chairman: Marshall J. Farr, Office of Naval Research

12:45 p.m. Discussants' presentations
Frederic M. Lord, Educational Testing Service
Bert F. Green, Johns Hopkins University
1:30 Discussion, presenters and discussants
2:15 Open Discussion
3:30 Closing Remarks Edmund F. Fuchs, Conference Coordinator

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

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

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

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

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

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

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

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

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

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

We notice that (2-6) is the reciprocal of the test information function, I(θ). 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

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

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

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

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 θ_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:

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

and

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

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	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 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-

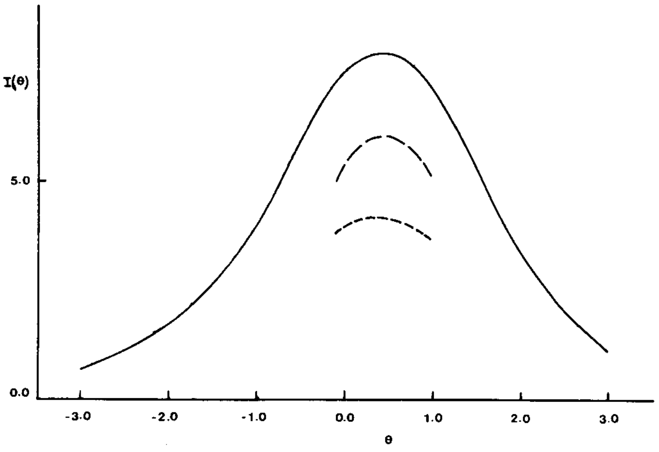


Figure 1. Test information functions for the subtest NMB, when the graded scoring strategy is taken (————), when the most informative dichotomous scoring strategy is taken for the interval [–0.1, 1.0] (– – – – –), and when the least informative dichotomous scoring strategy is taken for the interval [–0.1,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

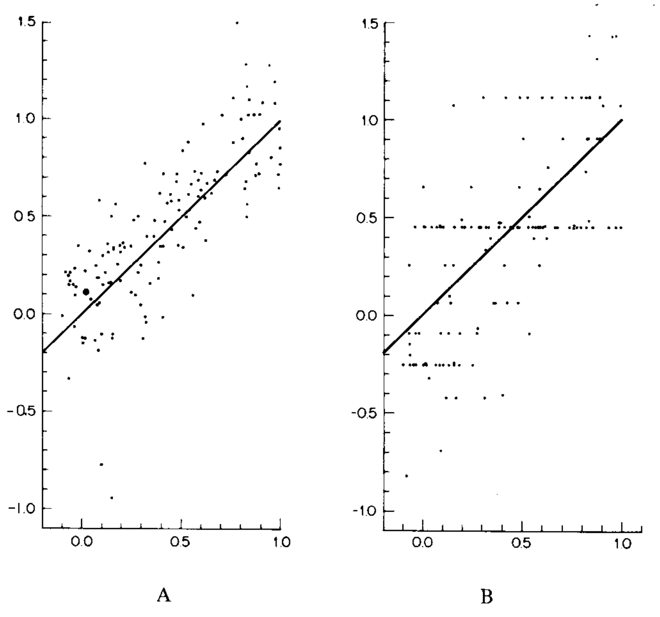


Figure 2. Maximum likelihood estimates obtained by dichotomizing NMB items for the interval [–0.1, 1.0], plotted against $\hat{\theta}$, those obtained from the original response patterns of graded item scores for the 138 subjects whose $\hat{\theta}$ are in the interval [–0.1, 1.0]. A. Using the most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

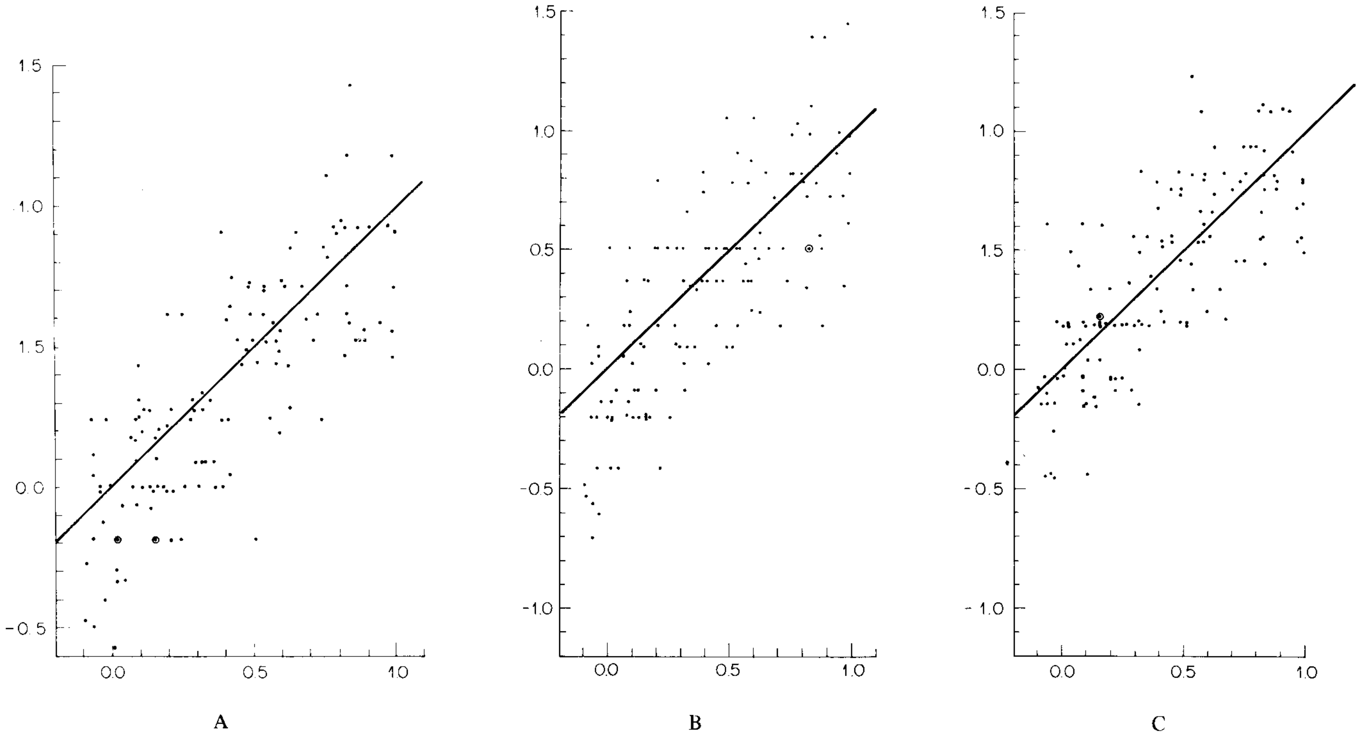


Figure 3. Maximum likelihood estimates obtained by simulated tailored testing plotted against $\hat{\theta}$, those obtained from the original response patterns of graded item scores for the 138 subjects whose $\hat{\theta}$ are in the interval [−0.1, 1.0]: A. Using the most informative dichotomized 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-

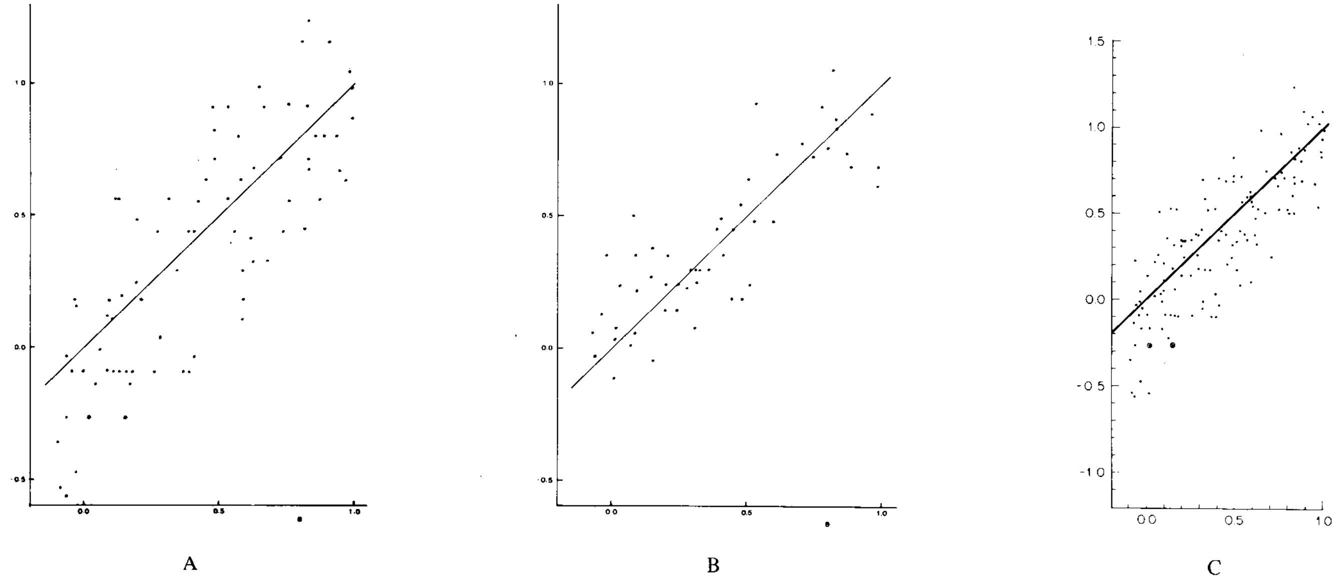


Figure 4. Maximum likelihood estimates obtained by simulated tailored testing plotted against c, those obtained from the original response patterns of graded item scores, for the subjects whose $\hat{\theta}$ are in the interval [–0.1, 1.0] : A. Using the most informative graded item, N7, as the initial 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,

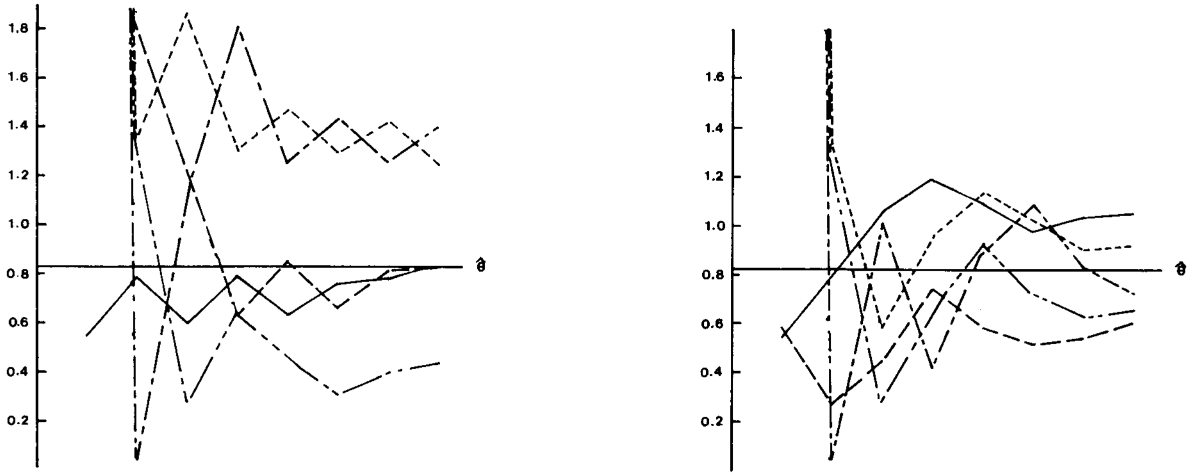


Figure 5. Two examples to show how the maximum likelihood estimates converge in the simulated tailored testing. Initial items are: N7, most informative graded item (————); N4, second most informative graded item (– – –); N7-2, most informative dichotomized item (- - - - -); N1-2, medium informative dichotomized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

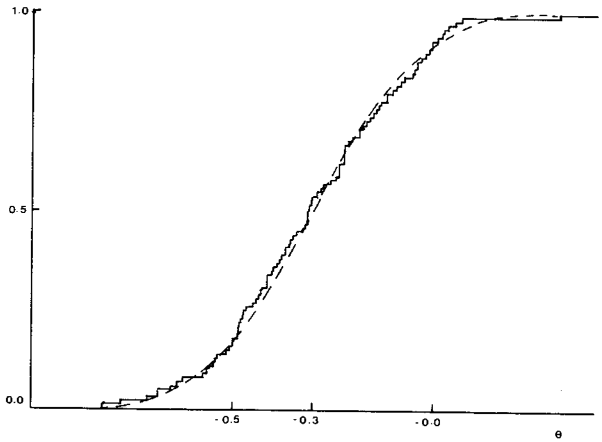
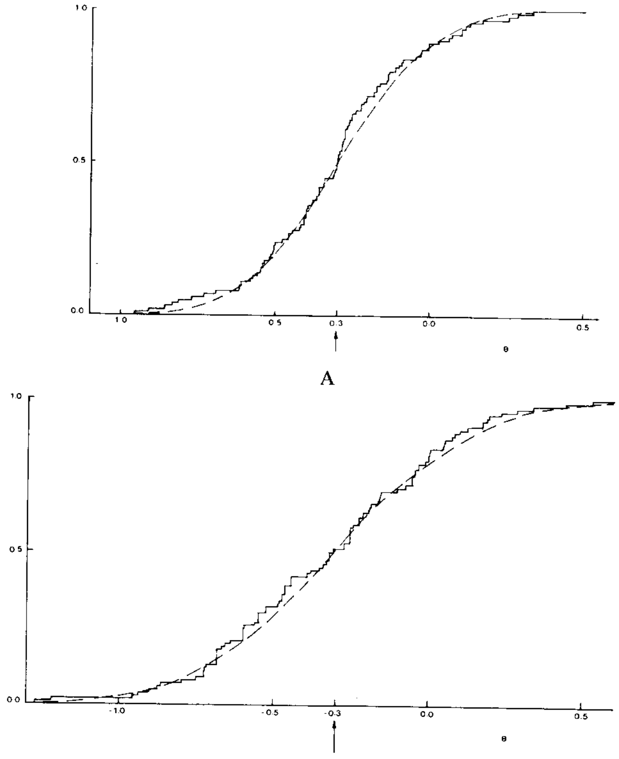


Figure 6. Cumulative frequency ratio of maximum likelihood estimates obtained from the original response patterns of graded item scores for the 100 hypothetical subjects (————) and the normal distribution function (---) with the parameters $\mu = -0.3$ and $\sigma = 0.2128$



A

B

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 $\sigma = 0.3685$ (– – –).

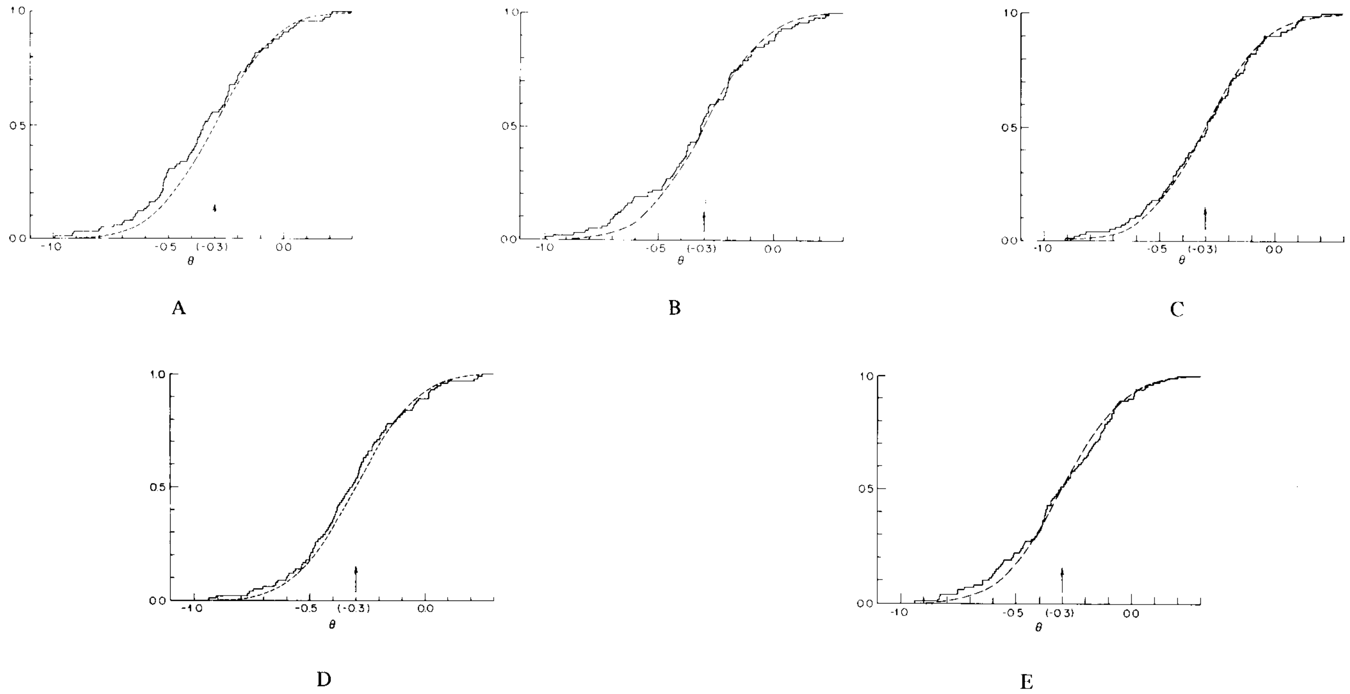


Figure 8. Cumulative frequency ratio of maximum likelihood estimates obtained by simulated tailored testing, for the 100 hypothetical subjects (– – –) and the normal distribution with the parameters $\mu = -0.3$ and $\sigma = 0.2128$ (– – –): A. with the most informative dichotomized item (23-2) as the 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	√MSE	1/MSE
	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
Dichotomous	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

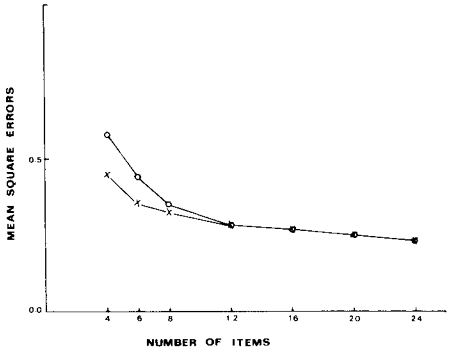


Figure 9. Comparison of the square roots of the mean square errors of the maximum likelihood estimates in simulated tailored testing with the graded item (24), plotted with *x* and the dichotomized item (14-3), plotted with *o*, as the initial item, calculated for 4, 6, 8, 12, 16, 20, and 24 items.

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.

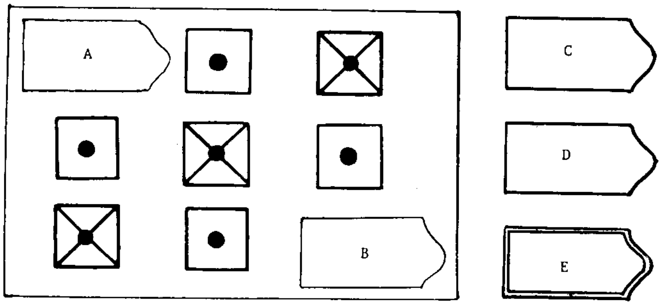
Tallis, L. R. The maximum likelihood estimation of correlation from contingency tables. *Biometrika,* 1962, 18, 342-353.

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: $\Omega e_1\ e_2 = (1 - 2)^2 \text{ x } (3 - 4)^2 = 1$

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 $\hat{\theta}$ Arc in the Interval [−0.8, 0.1].

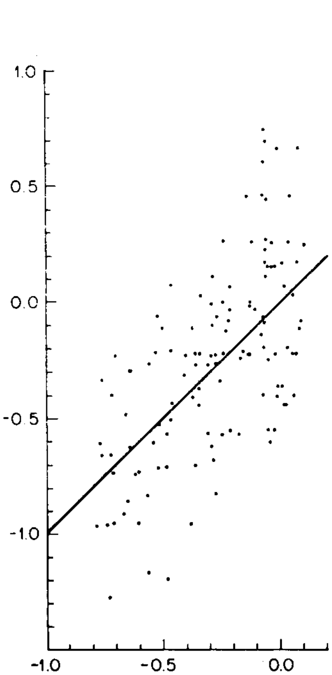


Figure B1

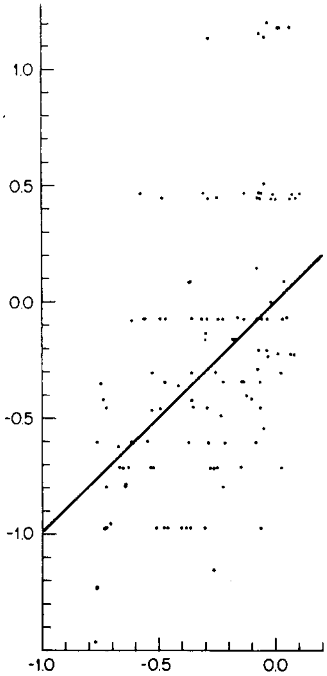


Figure B2

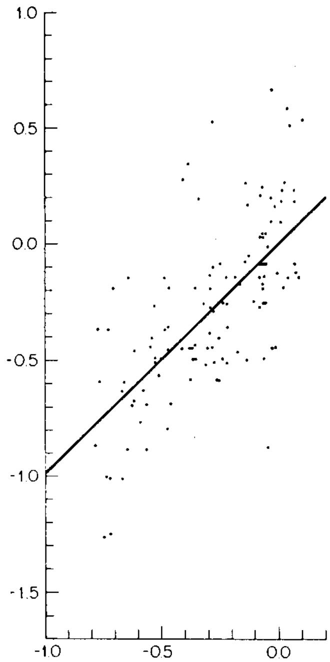


Figure B3

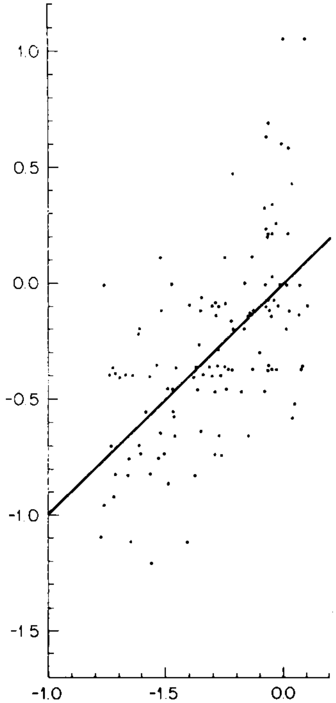


Figure B4

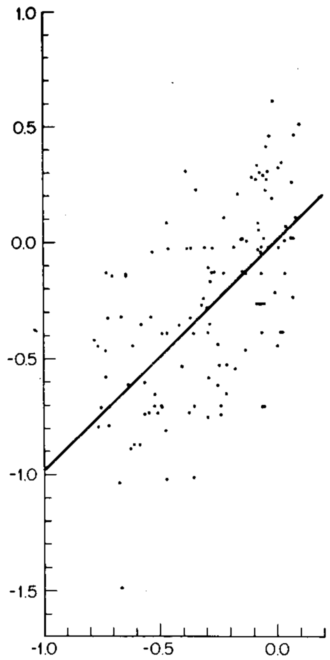


Figure B5

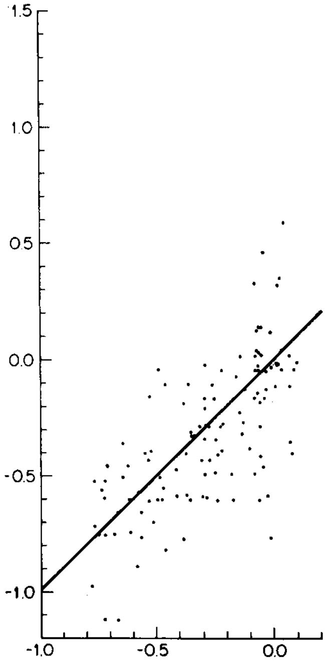


Figure B6

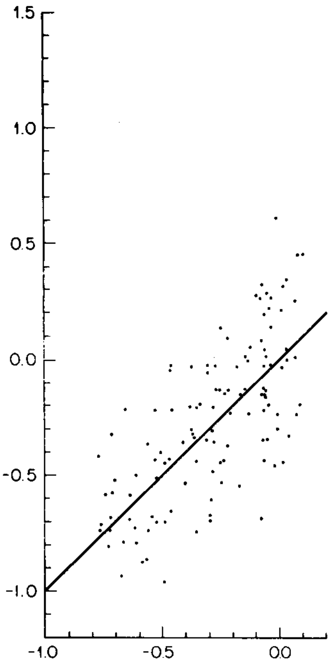


Figure B7

INCOMPLETE ORDERS AND COMPUTERIZED TESTING

NORMAN CLIFF[1]
University of Southern California

A computerized adaptive testing system has three main aspects, and consequently it can differ in three main ways from a noncomputer system. First, there is the test item. Full utilization of a computer allows an enormous broadening in the type of problem that can be presented to the individual. Typing out objective questions to him is the most obvious thing to do, but it is far from the only thing, and is perhaps far from the best thing. There is perhaps even a greater extension of the possible types of examinee response, as we can see not only from what is described here but by borrowing from CAI techniques. Moreover, we can easily incorporate speed of response into the scoring; we can determine not only whether the person can give the answer, but whether he can give it in ten seconds. But the greatest difference between computerized adaptive testing and ordinary testing is in the extent and nature of the decision process that goes on between items.

It is with the latter aspect that I will be concerned here today; the approach suggested here is quite different conceptually than others such as the branching and the Bayesian methods, so the paper will trace its origins. Tests try to order persons, so we will first consider the basic nature of orders and then how orders can be constructed from incomplete data. Testing will be shown to be a type of ordering process which utilizes incomplete data; computerized adaptive testing develops orders from highly incomplete data. We will give a simple example of how a computer program based on these concepts works. Finally, some of the ways in which these concepts form the basis for a test theory will be suggested.

Our approach to a model for computerized testing has its origins in quite a different area, computer-interactive judgment methods. In order to demonstrate the relation between testing and ordering, let us consider for a moment a simple order. A simple order is defined, and please let me use quite informal language, as a set whose members display a relation between elements which demonstrates asymmetry and transitivity. Now what that means is that, if we have a matrix which records the existence of the relation as a 1, or its non-existence as a 0, between a pair of elements of the set, the matrix must display the triangular form shown in the first figure. Paired comparisons judgments of some stimulus property of course often display a close approximation to this form. For example, suppose we used the five indicated letters, presented them in pairs, and asked a child which came first in the alphabet. Then we record his judgment as a 1 if he responds that the row letter comes before the column letter and a 0 if he says the reverse. If he knew the order of the alphabet, then the data would be as shown.

$$\begin{array}{cccccc}
 & v & w & x & y & z \\
v & - & 1 & 1 & 1 & 1 \\
w & 0 & - & 1 & 1 & 1 \\
x & 0 & 0 & - & 1 & 1 \\
y & 0 & 0 & 0 & - & 1 \\
z & 0 & 0 & 0 & 0 & -
\end{array}$$

Fig. 1. Complete adjacency matrix for a simple order showing transitivity and asymmetry.

An interesting property of such paired comparisons matrices is that they need not be complete. Suppose we do not ask about all pairs, but do assume that the data is asymmetric and transitive. Then we may be able to complete the matrix by performing matrix algebra on the elements which we do have. This is illustrated in the second set of figures. The lefthand one shows an incomplete dominance matrix, one which incidentally would typically be found by the kind of interactive ordering program we developed, and the right one shows that matrix multiplied by itself. We see that in this instance the square of the obtained matrix shows exactly the same triangular form as the complete matrix in Fig. 1. Actually, the data matrix could be even more incomplete than this one and still yield a complete order. The *necessary* part of the matrix is the supradiagonal chain of ones which corresponds to the judgments concerning the letters which are next to each other in the alphabet. As long as we have these, then the matrix can be completed; we just have to raise it to a high enough power. Of course, when dealing with human judgments with their inconsistency, we have to build in some safeguards and redundancy in the process.

The reason for going through that exercise is that the model we propose for computerized testing is exactly the

$$\begin{array}{cccccc}
 & v & w & x & y & z \\
v & - & 1 & & 1 & 1 \\
w & 0 & - & 1 & & 1 \\
x & & 0 & - & 1 & \\
y & 0 & & 0 & - & 1 \\
z & 0 & 0 & & 0 & -
\end{array}
\qquad
\begin{array}{cccccc}
 & v & w & x & y & z \\
v & - & & 1 & & 1 \\
w & & - & & 1 & \\
x & 0 & & - & & 1 \\
y & & 0 & & - & \\
z & 0 & & 0 & & -
\end{array}
\qquad
\begin{array}{cccccc}
 & v & w & x & y & z \\
v & - & 1 & 1 & 1 & 2 \\
w & 0 & - & 1 & 1 & 1 \\
x & 0 & 0 & - & 1 & 1 \\
y & 0 & 0 & 0 & - & 1 \\
z & 0 & 0 & 0 & 0 & -
\end{array}$$

Fig. 2. Sufficient adjacency matrix A_t, its square A_t^2 and the sum $A_t + A_t^2$, showing that the latter has the same qualitative form as A.

[1] Preparation of this paper was supported in part by the Office of Naval Research, Contract No. 150-373.

same! We say that tests order people. In what sense is that so? In what sense is the relation between people one which is asymmetric and transitive? It is superficially obvious that if examinees are given different scores, then the relation between the scores is asymmetric and transitive. That is just a property of numbers, in fact the one which served as a model for ordering in the first place. But it is a property which is just as true of the testees' zip-codes, or their social security numbers, or their football jersey numbers, as it is of their test scores. What is it about test scores that makes the order empirically meaningful rather than arbitrary?

Test scores start out from binary relations between people and items. How is it that we are allowed to derive from such relations numbers which give us an order of people, in the same sense that we can assign numbers to stimuli that give their order? Where is the asymmetric, transitive relation?

A long time ago, Louis Guttman gave part of the answer (Guttman, 1941). He said that items order persons if the score matrix displays the form we have come to call the Guttman scale, but should more fairly call the Guttman Loevinger scale since she invented an almost identical concept and developed it in a superior way (Loevinger, 1947). But Guttman's answer is not completely satisfactory to the formalist. The score matrix is rectangular, not square; item responses are defined as right or wrong by fiat and have no chance to be other then asymmetric. The transitivity of a Guttman scale is indirect.

The most important part of the answer to the questions concerning the legitimacy of items as orderers of persons lies in the realization that the score matrix is only part of a larger matrix of relations. The relations matrix is really items-plus-persons by items-plus-persons, not just items by persons. We think of the response of a person to an item as indicating a dominance relation between the person and the item. Habitually, we put a one in the score matrix if the person gets the item right and a zero if he gets it wrong. But that is because, being people, we identify with the persons dimension of the matrix. If instead we were items, in some through-the-looking-glass world, we would use the opposite notation, giving the *item* a one if the person got it wrong and a zero if the dumb thing allowed itself to be gotten right by the person.

$$\begin{array}{cccccc}
 & a & b & 1 & 2 & 3 \\
a & & & 0 & 1 & 1 \\
b & & & 0 & 0 & 1 \\
1 & 1 & 1 & & & \\
2 & 0 & 1 & & & \\
3 & 0 & 0 & & & \\
\end{array}$$

S

$$\begin{array}{cccccc}
 & a & b & 1 & 2 & 3 \\
a & 0 & 1 & & & \\
b & 0 & 0 & & & \\
1 & & & 0 & 1 & 2 \\
2 & & & 0 & 0 & 1 \\
3 & & & 0 & 0 & 0 \\
\end{array}$$

S^2

Fig. 3. Complete (showing rights and wrongs) score matrix S for two items a, b and three persons, 1, 2, 3 for scalable data; and S^2 showing item-item and person-person dominance.

Taking the point of view of neither items nor persons but rather of test theorists, we must take a less chauvinistic stance and play fair in our scorekeeping. The score matrix is expanded. In the expanded matrix, we give a one to the winner of the contest between item and person and a zero to the loser, regardless of which is which. Such a matrix is given at the left of Figure 3. In the lower left corner of the matrix we have the usual binary score matrix which shows which items were defeated by which persons. The matrix here is of the Guttman form. In the upper right we have the same matrix from the item point of view, giving a one each time an item defeats a person. Since the score matrix is complete here, the upper right matrix is the transposed complement of the lower right one.

There are two other sections of this expanded score matrix and these are left blank. These sections correspond to the item-item and person-person relations, which are not observed directly. In the case of pairwise judgments, we found above that an incomplete matrix could be completed by squaring the observed matrix. Let us do that in the present case. The result is shown in the right side of the figure. It is two triangular matrices, one for items and one for persons. Thus, treated in this formal fashion, we see that a GL scale does give two asymmetric transitive relations, one for items and one for persons. We will return to these two order matrices in another context.

We can put the two orders together. This is illustrated in Figure 4; the matrix on the left is simply the sum of the two matrices from Figure 3, that is $S + S^2$. The matrix on the right of Figure 4 contains exactly the same elements, but they have been rearranged, that is, pre- and postmultiplied by a permutation matrix P, into the order which is implied here, a *joint* order of persons and items, which is seen to in fact be a simple order because of the triangular, i.e., asymmetric and transitive form of the matrix. This answers those querulous questions about where the order is in the case of test data. If the data are a Guttman scale, then the score matrix, expanded and operated on in the manner indicated, does indeed define an order in the rather strict sense of the existence of a relation on a set, a relation which is transitive and asymmetric.

Let me say that for illustrative purposes here the matrix operations have been carried out in ordinary arithmetic.

$$\begin{array}{cccccc}
 & a & b & 1 & 2 & 3 \\
a & 0 & 1 & 0 & 1 & 1 \\
b & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 & 2 \\
2 & 0 & 1 & 0 & 0 & 1 \\
3 & 0 & 0 & 0 & 0 & 0 \\
\end{array}$$

$S + S^2$

$$\begin{array}{cccccc}
 & 1 & a & 2 & b & 3 \\
1 & - & 1 & 1 & 1 & 2 \\
a & 0 & - & 1 & 1 & 1 \\
2 & 0 & 0 & - & 1 & 1 \\
b & 0 & 0 & 0 & - & 1 \\
3 & 0 & 0 & 0 & 0 & - \\
\end{array}$$

$P(S + S^2)P'$

Fig. 4. $S + S^2$ in its original segregated form (left) and reordered form (right), the latter showing qualitative asymmetry and transitivity like a simple order.

Because the relations are logical rather than arithmetic, we should have been doing the matrix multiplication with Boolean arithmetic. The only thing that changes in the present context is that all numbers greater than one in the matrices should be set equal to one.

So far, we have not referred directly to anything having to do with "computerized adaptive testing," but the relevance of the above theoretical sketch is quite direct. Just as the score matrix itself is a kind of incomplete matrix of dominance relations that can be completed by the powering operation, an even more incomplete set of relations is all that is really necessary to define the joint person-item order. If we happen to ask each person only the hardest item he can answer correctly and the easiest item he would miss, those 2n relations—actually, 2n-2 is enough—are sufficient to define the complete joint order of items and persons. This subset of relations can quite simply be shown to correspond to the relations between adjacent elements in the order, the supradiagonal string of ones we saw in the incomplete paired comparisons matrix of Fig. 2. In fact, if you look at the righthand matrix of Figure 4, the string of ones just above the diagonal there denotes exactly this set of item-person relations. In the 1975 *Bulletin* article (Cliff, 1975) I illustrated the way in which such a set of

$$\begin{array}{cccccccccc}
 & a & b & c & d & 1 & 2 & 3 & 4 & 5 \\
a & & & & & 0 & 1 & 0^* & 0^* & 0^* \\
b & & & & & 0 & 0 & 1 & 0^* & 0^* \\
c & & 0 & & & 0^* & 0 & 0 & 1 & 0^* \\
d & & & & & 0^* & 0^* & 0 & 0 & 1 \\
1 & 1 & 1 & & & & & & & \\
2 & 0 & 1 & 1 & & & 0 & & & \\
3 & & 0 & 1 & 1 & & & & & \\
4 & & & 0 & 1 & & & & & \\
5 & & & & 0 & & & & &
\end{array}$$

A

$$\begin{array}{cccccccccc}
 & a & b & c & d & 1 & 2 & 3 & 4 & 5 \\
a & & & & & 0 & 1 & 1^* & 1^* & \\
b & & & & & 0 & 0 & 1 & 1^* & 1^* \\
c & & & 0 & & 0^* & 0 & 0 & 1 & 1^* \\
d & & & & & 0^* & 0^* & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 0^* & & & & & \\
2 & 0 & 1 & 1 & 1^* & & 0 & & & \\
3 & 0^* & 0 & 1 & 1 & & & & & \\
4 & 0^* & 0^* & 0 & 1 & & & & & \\
5 & & 0^* & 0^* & 0 & & & & &
\end{array}$$

$A(A + I)^{(2)}$

$$\begin{array}{cccccccccc}
 & a & b & c & d & 1 & 2 & 3 & 4 & 5 \\
a & 0 & 1 & 1 & 1 & & & & & \\
b & 0 & 0 & 1 & 1 & & 0 & & & \\
c & 0 & 0 & 0 & 1 & & & & & \\
d & 0 & 0 & 0 & 0 & & & & & \\
1 & & & & & 0 & 1 & 1 & 1 & 0 \\
2 & & & & & 0 & 0 & 1 & 1 & 1 \\
3 & & 0 & & & 0 & 0 & 0 & 1 & 1 \\
4 & & & & & 0 & 0 & 0 & 0 & 1 \\
5 & & & & & 0 & 0 & 0 & 0 & 0
\end{array}$$

$A(A + I)^{(3)}$

$$\begin{array}{cccccccccc}
 & a & b & c & d & 1 & 2 & 3 & 4 & 5 \\
a & & & & & 0 & 1 & 1^* & 1^* & 1^* \\
b & & & 0 & & 0 & 0 & 1 & 1^* & 1^* \\
c & & & & & 0 & 0 & 0 & 1 & 1^* \\
d & & & & & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 1^* & 1^* & & & & & \\
2 & 0 & 1 & 1 & 1^* & & & & & \\
3 & 0^* & 0 & 1 & 1 & & 0 & & & \\
4 & 0^* & 0^* & 0 & 1 & & & & & \\
5 & 0^* & 0^* & 0^* & 0 & & & & &
\end{array}$$

$A(A + I)^{(4)}$

Fig. 5. Illustration of completion by powering. Starred entries are derived by implication.

relations could be used to reconstruct the complete score matrix. That process is reproduced here in Figure 5 where the matrix powering is carried out.

Unfortunately, there is a problem; we do not know the right items to ask a person until after we have asked them. The routine by which the computer searches for the right items to ask is one of the two main aspects of the processing part of computerized adaptive testing, the other main aspect being how it damps out error. In our research, what we are doing is carrying over some principles which we have previously found to be effective in the paired comparisons ordering case.

The next set of figures illustrate the operation of a prototype program of the kind we have in mind, written by Jerry Kehoe. First, the program asks each person two items at random. The entries in the lefthand matrix of Figure 6 show the results of these preliminary rounds and the righthand one shows the powered matrix which contains the implications of these responses as well as the responses themselves. So far these are very few. The computer then decides which items to ask which persons next by seeing which are closest together in the order so far determined. This process of presentation, powering, and selection would go on for several rounds. The next figure shows the score matrix for an intermediate round on the left and the implications on the right. Now the powering process is having some effect. The next one shows the final score matrix on the left and the implications on the right where we see that not only has the score matrix been completed by implication but there are now complete simple orders of persons and items.

We incidentally do not have a name for this method. We would like to call it the Extended Transitivity System, or ETS, but those initials have been preempted.

You can see that the savings are not very great in this instance; each person must be asked most of the items. This impression is primarily a function of the size of the data matrix here. The savings are much, much greater with large matrices. An upper bound for the number of item-person relations that must be observed for *n* persons and *x* items is $\log_2(n + x)!$. For 200 persons and 200 items this number is about 2886. That means we would need to ask each person only 15 items to get the complete order; moreover, this upper bound is quite a generous one in the present instance, a couple fewer might well be sufficient.

Thus the method will work if the responses form a Guttman scale. It works surpiisingly quickly and requires surprisingly little space in the computer, primarily because the programs take advantage of the binary nature of the data to store responses as single bits and then to carry out many of the calculations on whole words, that is, 32 elements at a time are processed in raising the matrix to the next power.

It is really no surprise that it works with errorless data. The crucial questions are how well will it work with the kind of inconsistent items and persons that the real world

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a													
b								0					
c								0	0				
d											1		
e									0		0		
f										0		0	1
g										0		0	1
1		1	1										
2			1		1								
3						1	1						
4				0	1								
5						1	1						
6						0	0						

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a													
b								0					
c								0	0				
d					1						1		
e				0					0		0		
f										0		0	1
g										0		0	1
1		1	1										1
2			1		1								
3						1	1						
4				0	1								
5						1	1						
6						0	0	0					

Fig. 6. (Left) Initial item responses matrix S, showing both person dominances and item dominances. Blank entries indicate item-person pairs not yet observed. (Right) S + S², showing the implied item-item and person-person dominances.

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a								1					1
b								0		1			1
c								0	0		1		
d									0		1		
e									0		0	1	
f										0		0	
g										0		0	1
1	0	1	1										
2			1	1	1								
3		0				1	1						
4			0	0	1								
5					0	1	1						
6	0	0				0							

	items							persons					
	a	b	c	d	e	f	g	1	2	3	4	5	6
a		1	1		1*	1*	1*	1		1*	1*	1*	1
b	0					1	1	0		1			1
c	0				1	1*	1*	0	0		1	1*	1*
d					1	1*	1*		0		1	1*	1*
e	0*			0	0	1	1	0*	0		0	1	1*
f	0*	0	0*	0*	0			0*	0*	0	0*	0	1
g	0*	0	0*	0*	0			0*	0*	0	0*	0	
1	0	1	1		1*	1*	1*			1	1	1*	1
2			1	1	1*	1*	1*			1	1	1*	1*
3	0*	0				1	1	0	0				1
4	0*		0	0	1	1*	1*	0	0			1	1*
5	0*		0*	0*	0	1	1	0*	0*		0		1
6	0	0	0*	0*	0*	0		0	0*	0	0*	0	

Fig. 7. (Left) intermediate item response matrix S. (Right) $S + S^{(2)} + S^{(3)} + S^{(4)} + S^{(5)}$. Starred (*) entries are derived by indirect implication, i.e., from $S^{(3)}$, $S^{(4)}$, or $S^{(5)}$.

	a	b	c	d	e	f	g	1	2	3	4	5	6
a								1	1				1
b								0	1	1	1		1
c								0	0	1	1		
d								0	0	0	1		
e									0	0	0	1	
f										0		0	1
g										0		0	0
1	0	1	1	1									
2	0	0	1	1	1								
3		0	0	1	1	1	1						
4		0	0	0	1								
5					0	1	1						
6	0	0				0	1						

	a	b	c	d	e	f	g	1	2	3	4	5	6
a		1	1	1	1	1*	1	1	1	1*	1*	1*	1
b	0		1	1	1	1	1	0	1	1	1	1*	1
c	0	0		1	1	1	1	0	0	1	1	1*	1*
d	0	0	0		1	1*	1*	0	0	0	1	1*	1*
e	0	0	0	0		1	1	0*	0	0	0	1	1*
f	0*	0	0	0*	0		1	0*	0*	0	0*	0	1
g	0	0	0	0*	0	0		0*	0*	0	0*	0	0
1	0	1	1	1	1*	1*	1*		1	1	1	1*	1
2	0	0	1	1	1	1*	1*	0		1	1	1	1*
3	0*	0	0	1	1	1	1	0	0		1	1	1
4	0*	0	0	0	1	1*	1*	0	0	0		1	1*
5	0*	0*	0*	0*	0	1	1	0*	0	0	0		1
6	0	0	0*	0*	0*	0	1	0	0*	0	0*	0	

Fig. 8. (Left) Final response matrix S, showing 26 of the 42 item-person combinations which were used. (Right) $S + S^{(2)} + S^{(3)} + S^{(4)} + S^{(5)}$ with starred (*) elements indicating those entered by indirect implication.

faces us with, and what advantages does it offer over other approaches? The answer to the first question must await the opportunity to test it first with artificial stochastic data and then with real data. How well it will do in practice relative to the other approaches that have been reported and which we are hearing about during these two days must await even further data.

A priori, the methodology here appears to offer at least one potential advantage, the avoidance of extensive pretesting to determine item characteristics. Such pretesting presented problems, even to paper and pencil testing. There was the security problem, the question of comparability of populations, the differing contexts, the expense itself. In the computerized situation, these all become more acute. The present process avoids pretesting since items and persons are processed in parallel.

This method does require a substantial number of persons being tested simultaneously, however; but this is only initially true. Once a substantial set of person-item relations has been built up, additional persons can be processed individually as they appear, being fit into the previously determined order by means of their responses to the items. Under that mode of operation the amount of additional computer processing would be quite small.

It also seems to me that this way of thinking about tailored testing makes it easier to think of testing as integrated into a total personnel process. After all, it could be that the item selected for a person at a given point could be something like, "You have been assigned to welders' school. Come back when you have completed the course." The "item" in that case is successful completion of the course.

But to me, the most promising aspect of this method is theoretical. It furnishes the basis for a test theory which I think is more appropriate to the computerized testing context. If what is wanted from testing is an order of persons, and norms after all just tell the individuals' positions relative to some benchmark persons, then surely we want the order to be consistent and complete. How do you tell if the order is consistent and complete? You look at the person-person relation matrix and see if it is asymmetric and transitive. It is easy to think of indices which would reflect the degree to which that matrix has those properties. Indeed, I had intended to spend my time here today talking about them, but the results of our study are not quite ready for presentation yet. Such indices furnish analogues of the familiar Kuder-Richardson formulas which are central to basic test theory, and in fact are related to them in the case of complete data. They have the additional property of being readily generalizable to the incomplete or computer-adaptive case. Thus if we go about computerized testing in the way described here, we can at least have appropriate evaluational indices built into the system. Other tailored testing schemes rely on external information from traditional modes of testing to get their biserial correlations, item difficulties, reliabilities, and so on. Here, analogues of these indices will come out of the interactive process itself.

REFERENCES

Cliff, N. Complete orders from incomplete data: Interactive ordering and tailored testing. *Psychological Bulletin,* 1975,*82,* 289-302.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Loevinger, J. A systematic approach to the construction and evaluation of Tests of Ability. *Psychological Monographs,* 1947, 61 (4, Whole No. 285).

ADAPTIVE TESTING RESEARCH AT MINNESOTA—OVERVIEW, RECENT RESULTS AND FUTURE DIRECTIONS [1]

DAVID J. WEISS
University of Minnesota

Adaptive Testing and Error Reduction

The general objective of our research program on adaptive testing is to view it from a perspective which identifies several sources of potential error in test scores, and to study adaptive testing as a means for reducing these errors of measurement.

The first general source of error that we have been concerned with for some time is the error that results from the mismatch of item difficulties in an ability test with the individual's ability. Obviously, the testee's ability is not known at the start of testing. But the different strategies of adaptive testing that have been proposed can be viewed as different ways of matching item difficulties with testee ability and sequentially estimating the testee's ability. Consequently, one of our major focuses is to determine the best, or at least better, ways of adapting item difficulties to individual abilities.

We are approaching this in two complementary ways. First, we have been doing live computerized testing. Since late 1972 we have tested more than 5,000 subjects on a variety of strategies of adaptive testing. But live testing cannot provide the answer to all the questions concerning which strategies are best under which conditions, because there are too many questions to be answered. Therefore, we are using computer simulation to supplement and extend the results that we obtain from live testing.

Our general strategy is to implement an adaptive testing strategy in live testing to obtain some data with an arbitarily structured live adaptive test—data such as characteristics of score distributions and test-retest reliabilities. Then, our ultimate goal is to build a computer simulation model which will accurately reflect the results that we obtain from live testing. With the computer simulation model we can then very rapidly study different variations of the adaptive testing strategy. The next step is to verify the simulation results in live testing.

Thus far we have not yet developed a simulation model which completely reflects how live testees respond, but we are making progress toward that goal. The computer simulations are necessary because of the rapidity with which we can study various alternatives. The live testing is necessary, obviously, because it's people who take tests and not computers using hypothetical items or hypothetical subjects. So it is necessary to re-verify the results of the computer simulations to make sure that they still reflect what real people do given the variations we have made in the strategies studied in the simulations.

The second main focus of our research is a concern with the psychological effects of adaptive testing. Here we are concerned with identifying the psychological aspects of testing and the test environment which can introduce error into test scores. These variables include guessing, test anxiety, boredom, frustration, and racial or ethnic group effects.

Guessing can obviously artifically increase test scores; frustration, anxiety, motivation and other factors can result in test scores lower than true ability. All of these, therefore, are sources of error in test scores which are due to the psychological effects of testing.

We are also concerned with the psychological effects that will result from the man-machine interface. This, from our experience, is going to be an important problem in computerized adaptive testing. There are different kinds of computer systems on which we can implement adaptive testing and each of those computer systems has its positive and negative effects on testee behavior. There are different kinds of terminal devices for adaptive testing and each kind of terminal device displays in different ways and at different speeds. All of these variations in the man-machine interface are going to be new problems for us to consider in the years to come. Past research has demonstrated that answer sheets in paper and pencil testing sometimes had an effect on test scores. Similarly, research in adaptive testing will need to study different kinds of CRTs, different kinds of computer systems and different display speeds as part of the psychological effects of computerized testing.

A third source of error that we are concerned with has been briefly discussed this morning by Dr. Samejima; this is error that results from not extracting enough information from a testee's response to a test item. To date, most psychometric research has been concerned with binary or 0-1 scoring. But, as Dr. Samejima has indicated, we can get more information out of a test response if we treat it as a graded item. Our research extends that reasoning to continuous responses using the continuous case of latent trait theory. The continuous case is operationalized by probabilistic responding.

[1] Early development work on this research was supported during 1969 and 1970 by grants from the General Research Fund of the Graduate School, University of Minnesota. Research reported in this paper was supported since early 1972 by Personnel and Training Research Programs, Office of Naval Research, Contract No. N00014-67-A-0113-0029, NR 150-343. Special thanks are due to John DeWitt, our project programmer, without whom this research would have been almost impossible.

This aspect of our research is concerned with integrating probabilistic responding with adaptive testing. Probabilistic responding, like adaptive testing, can result in horizontal information functions. This implies that if we put adaptive testing and probabilistic responding together we will have extremely powerful methods of reducing errors in test scores due to the incomplete use of test responses.

The fourth source of error that we are studying is the error that results from deviations from unidimensionality. Latent trait theory, as it is usually used in testing, is based on the assumption of unidimensionality, although there are multidimensional latent trait models being developed. But dimensionality that is defined on a group, such as the unidimensionality of latent trait theory, does not necessarily hold true for an individual. That is dimensionality defined by factor analysis or other methods, when applied to an individual, assumes that the individual is the typical or average member of the group on which the dimensionality was defined. Thus, in the testing situation, when a set of "unidimensional" items is administered to an individual, the result may be a set of responses that are not unidimensionally determined.

Consequently, our research is concerned with individual-item pool interactions—the interaction of one individual with a set of "unidimensional" items. We are studying item response protocols of this nature to determine if meaningful deviations from unidimensionality do occur for specific individuals. If they do, we will then attempt to develop interactive testing models that will take account of intra-individual multidimensionality in an adaptive testing situation.

The focus of our research effort, as you can see, is with the *individual.* We are concerned with identifying those sources of error in test scores which result in the over- or under-estimation of *each individual's* ability.

Recent Results

Most of our recent results are concerned with the psychometric effects of adaptive testing, or the comparison of branching strategies. Thus far we have reported initial results from both live testing and computer simulation on a simple two-stage test (Betz & Weiss, 1973, 1974; Larkin & Weiss, 1975) and a pyramidal branching strategy (Larkin & Weiss, 1974, 1975). Below, I will report some results from a flexilevel test (Betz & Weiss, 1975) and some data on my stratified adaptive test (Weiss, 1975). Mr. McBride will present some data using Owen's (1975) Bayesian adaptive testing strategy.

In general, the findings that we have to date show that adaptive tests have higher test-retest stabilities—a very practical and useful criterion—when controlled for number of items and memory effects. Adaptive tests also tend to show, in simulation studies, better distributions of ability estimates. That is, ability estimates better reflect the distribution of generated ability. And, in general, adaptive tests give information functions which are less variable throughout the ability range, in support of Lord's theoretical findings (see Weiss & Betz, 1973).

Flexilevel ability testing. Figure 1 shows the item structure for Lord's (1971a,b) flexilevel test. In this testing strategy there is one item at each of a number of difficulty levels; item 19 is the most difficult item and item 18 the least difficult item. Everyone starts the flexilevel test with an item of median difficulty. Items with odd numbers increase in difficulty as they deviate from the median, and items with even numbers decrease in difficulty.

Figure 2 shows the paths taken by three different people through a ten-stage flexilevel test. Starting with the first item, a correct response leads to the next more difficult item which has not yet been administered. An incorrect response leads to the most difficult of the unadministered easier items. Figure 2a shows a high ability testee going through a flexilevel test, Figure 2b is for an average ability testee, and Figure 2c is for a low ability testee.

Our live-testing study of flexilevel testing (Betz & Weiss, 1975) used a flexilevel test in which each testee would answer 40 items, requiring a 79-item structure. That test and a conventional peaked paper-and-pencil type test, administered on a computer to control for novelty effects, was administered to 130 individuals. The same tests were then used in a computer simulation study. That study used 10,000 "subjects" sampled from a normal distribution of ability, and an additional 1600 subjects, 100 at each of 16 levels of ability. From these simulation data we calculated information functions, and test-retest or parallel forms reliability. From the live-testing study we calculated test-retest reliabilities, and other data describing score distributions.

The major result from the live-testing study was that flexilevel test scores were no more stable on retest than scores on the conventional test; test-retest stabilities for the two were virtually identical. The major result from the simulation study is shown in Figure 3, which displays information functions for the conventional and flexilevel tests. Figure 3 shows two findings which were not predicted by test theory.

First, test theory (e.g., Lord, 1971c) predicts that the conventional test will always result in higher levels of information, i.e., better measurement, than any adaptive test at the median of the ability distribution. Figure 3 shows that the flexilevel test had higher levels of the information function at the median (θ=0) of the ability distribution. The second prediction from test theory (Lord, 1971b) was that the flexilevel test should yield a relatively horizontal information function. Figure 3 shows an information function for the flexilevel test which is quite divergent from horizontal. In fact, the standard deviations of the information functions show that the flexilevel test had a larger standard deviation than did the conventional test; that means that the flexilevel test tended to be less equi-precise than the conventional test, at different levels of the ability distribution.

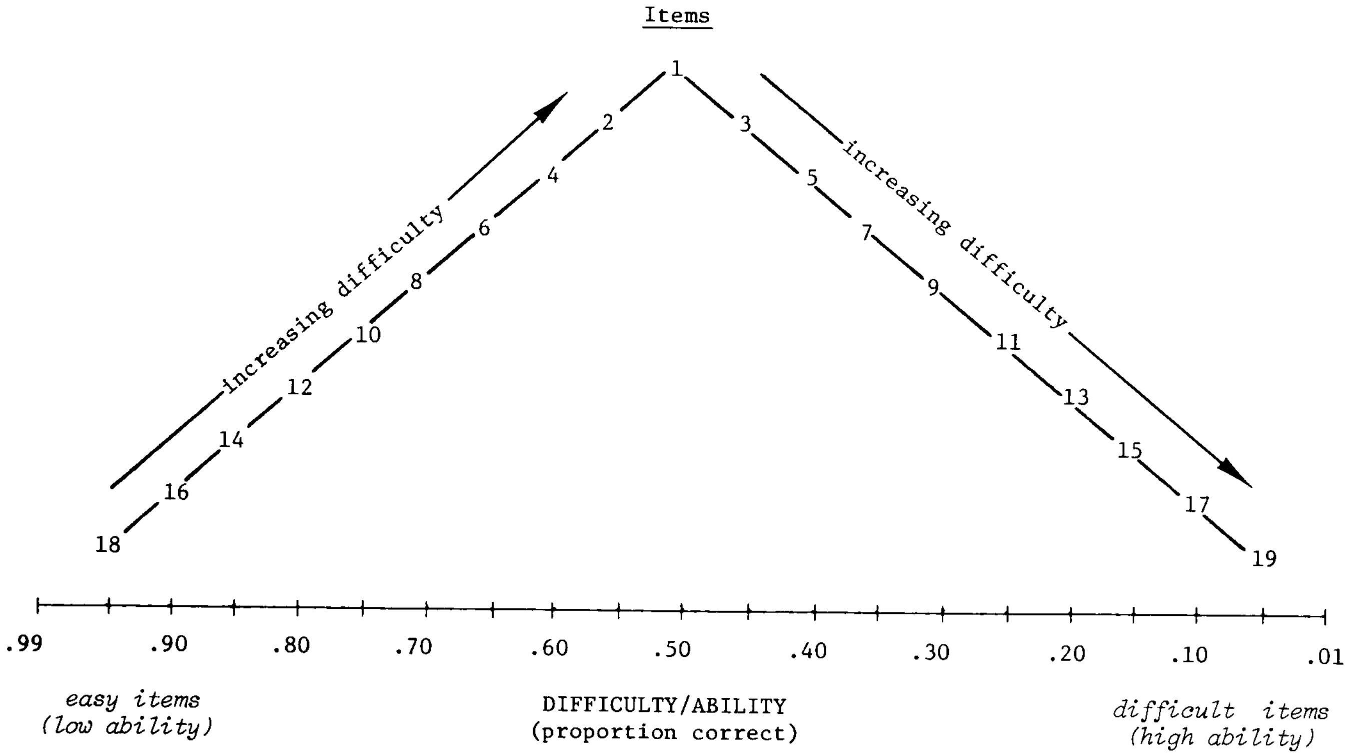


Figure 1

Items Structure for a Ten-stage Flexilevel Test

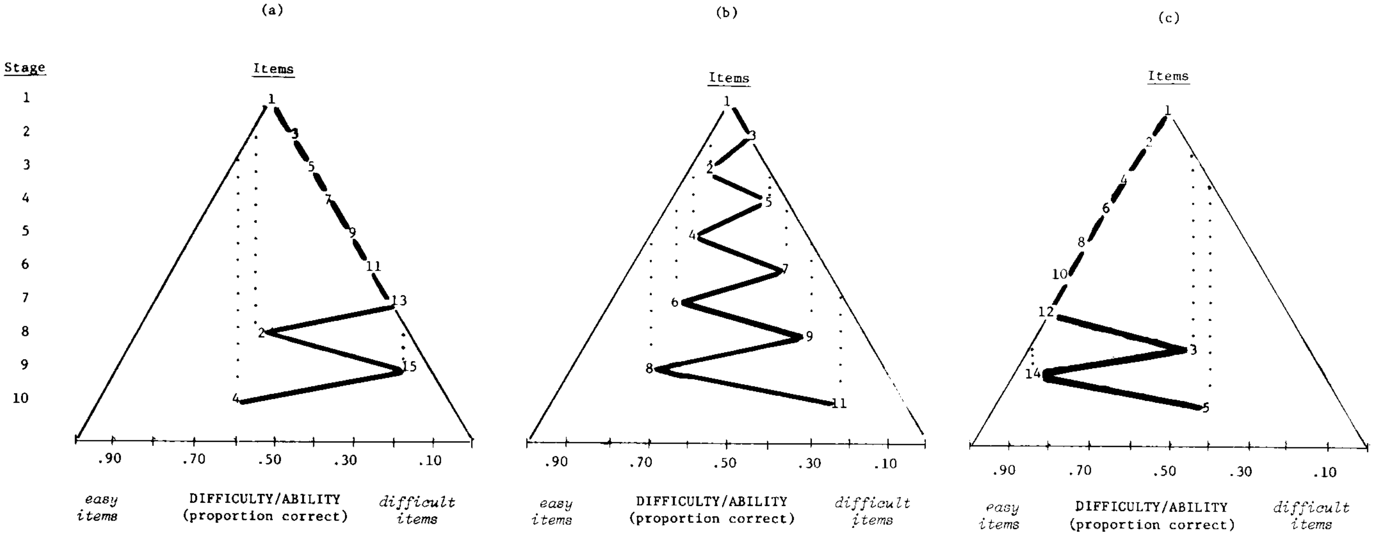


Figure 2

Sample Paths through a Ten-stage Flexilevel Test

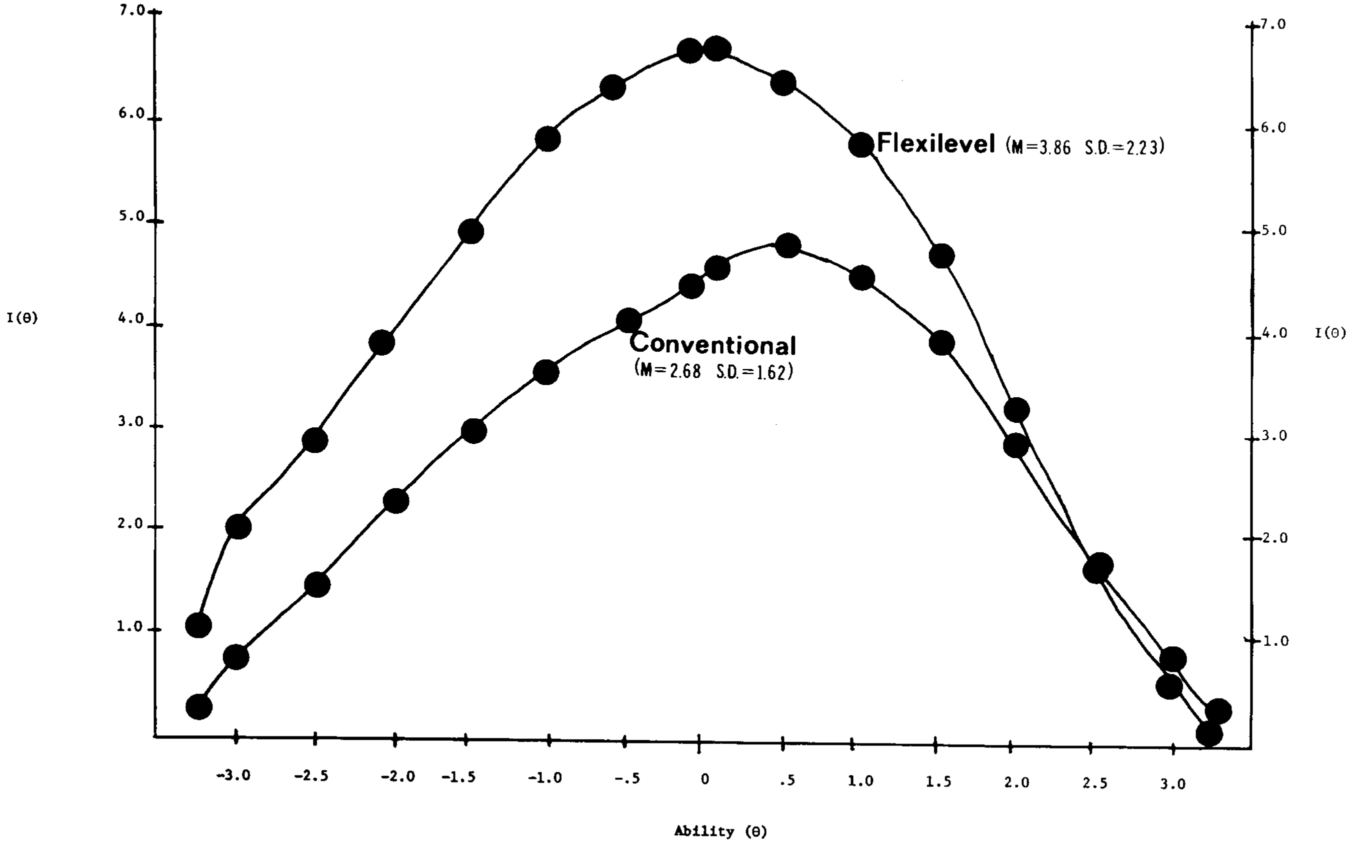


Figure 3

Information functions from flexilevel and conventional tests
(N=100 at each of 16 levels of ability)

A comparison of the results from the computer simulation study and the live-testing study showed differences in the test-retest reliabilities. This result was expected because of the memory effects in live testing. There were also differences between the two studies in the shapes of the generated score distributions. These differences demonstrated that the simulation model was not yet adequate enough to reflect the results of live testing and that it needs some revision so that it will enable us to extrapolate from live testing through computer simulation and back to live testing.

Another interesting result from this simulation study relates to the methodology of computer simulation itself. The design of the study was one in which we repeated the computations for a hundred samples of a hundred subjects each in order to study the sampling distribution of the simulation results. This was done to examine the generality of findings from computer simulation studies which use 100 or fewer simulated subjects (e.g., Jensema, 1974; Urry, 1971). We found that estimates of validity, the correlation of generated ability with estimated ability, based on samples of 100, ranged from .87 to .95, with a mean of .91. In certain inter-strategy comparisons different conclusions about the relative utility of a testing strategy might be drawn based on validities of .87 or .95. Thus, simulation studies should be based on samples of more than 100 in order to arrive at stable conclusions.

Two-stage testing. Figure 4 shows a computer report from what we have called a continuous second-stage two-stage test. This adaptive testing procedure was developed by Brad Sympson of our research staff; we later discovered that Fred Lord had independently developed the same testing procedure. In Fall 1975 we tested a number of college students on this continuous second-stage test.

The major problem with two-stage tests as they have been used in the past (Weiss, 1974) is that of routing errors made in branching from the routing test to the measurement test because of errors of measurement in the routing test. To solve this problem, we developed a measurement test stage which consists of a number of very short measurement tests. The example shown in Figure 4 used a 14-item routing test and 25 4-item measurement tests, each at a different level of difficulty. Using this adaptive testing procedure, when an individual completes

```
                                   REPORT ON TWO - STAGE TEST

NAME:                              DATE TESTED: 74/12/ 3                                   I.D. NUMBER:


                                     14 ITEM ROUTING TEST

                     X = CORRECT    0 = INCORRECT    . = ITEM NOT ADMINISTERED

   I-----------------------------------------------------------------------------------------I
   I                                                                                         I
   I    X     X     X     X     X     X     X     X     X     X     0     0     0     0     I
   I                                                                                         I
   I-----------------------------------------------------------------------------------------I
   (EASY)                 > > > > > > INCREASING DIFFICULTY > > > > > >                  (HARD)

   ESTIMATED ABILITY FROM ROUTING TEST:  1.4                    MEASUREMENT TEST ENTRY POINT: 18



                                  100 MEASUREMENT TEST ITEMS
                                 (36 ITEMS ARE ADMINISTERED)

      DIFFICULTY LEVELS:

      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25

   I-----------------------------------------------------------------------------------------------------I
   I                                                                                                     I
   I  .   .   .   .   .   .   .   .   .   .   .   .   .   X   0   X   X   0   X   X   X   0   .   .   .  I
   I                                                                                                     I
   I  .   .   .   .   .   .   .   .   .   .   .   .   .   X   X   X   0   X   X   X   0   0   .   .   .  I
   I                                                                                                     I
   I  .   .   .   .   .   .   .   .   .   .   .   .   .   X   X   X   X   X   0   X   X   0   .   .   .  I
   I                                                                                                     I
   I  .   .   .   .   .   .   .   .   .   .   .   .   .   0   X   X   X   X   X   X   0   X   .   .   .  I
   I                                                                                                     I
   I-----------------------------------------------------------------------------------------------------I

     -3.0     -2.5     -2.0     -1.5     -1.0     -.5      0       .5      1.0      1.5      2.0      2.5      3.0

   (EASY)                                                                                               (HARD)


                                  FINAL ABILITY ESTIMATE:   1.30

                                  PERCENTILE SCORE      :  90.26


OTE: NORM GROUP FOR ABILITY ESTIMATE AND PERCENTILE SCORE CONSISTED OF
     UNIVERSITY OF MINNESOTA UNDERGRADUATE AND GRADUATE STUDENTS
```

Figure 4

REPORT ON TWO-STAGE TEST

the routing test his score is determined and that score is used to choose an appropriate measurement test. Then, to reduce routing errors, a number of measurement tests on either side of the chosen measurement test are also administered to the individual. In the example shown in Figure 4, the individual's score on the routing test estimated his ability at 1.4 standard deviations above the mean. Consequently, the most appropriate measurement test was estimated to be number 18, which had items at difficulty about 1.4 standard deviations above the mean. But, to compensate for possible errors of measurement in the routing test, he was also administered items in measurement tests 14 through 17 and 19 through 22, for a total of 36 measurement test items. These items varied in difficulty from about .25 to 2.25 S.D.'s on the difficulty continuum.

Following a design that we have used in a number of other studies, we did a test-retest live-testing study with this continuous second-stage two-stage test (in which each testee completed 50 items) and a 50-item conventional peaked test, over about a five-week period, with 104 testees. To keep scoring method the same for both testing strategies, maximum likelihood scoring was used for both the two-stage and conventional test.

The study was designed also to equate the two testing procedures for 1) item discriminations; 2) memory effects; and 3) number of items. Memory effects were equated by first determining the number of items each individual repeated on retest of the two-stage test. Then the retest of the conventional test was structured to have the same number of repeated items by inserting the appropriate number of new items.

The test-retest correlation was .94 for the continuous two-stage test and .66 for the equivalent conventional test. Since the difference in stabilities was considerably larger than found in our previous studies of conventional vs. adaptive testing strategies (e.g., Betz & Weiss, 1973, 1975; Larkin & Weiss, 1974), we carefully examined the distribution of conventional test scores derived from the maximum likelihood scoring. Six testees were found with very low ability scores, apparently due to guessing on the conventional test. Data for these testees were eliminated and the test-retest correlations were recalculated. The stability correlation for the two-stage test was .93 and the conventional test .89. This result was similar to that obtained in other comparisons of conventional and adaptive strategies, showing a higher test-retest correlation for the adaptive test than for the peaked conventional test. This result was obtained when both testing strategies were equated for item discriminations and memory effects.

Stradaptive ability testing. The stradaptive testing strategy (Weiss, 1973) is based on a series of peaked tests, each one differing in terms of difficulty. Figure 5 shows the distribution of item difficulties for a hypothetical stradaptive test. In Figure 5 there are nine strata, each of which is a peaked test peaked at a different level of difficulty.

Figure 6 shows an example of an individual moving through a stradaptive test. Testing begins with an item at some point on the difficulty continuum; the entry point is estimated by prior information about the testee. The individual shown in Figure 6 began with the first item at stratum 5, an item of average difficulty. Since he answered that item correctly, he was administered the first item at stratum 6, which consisted of slightly more difficult items. Following the same branching rule—a more difficult item is administered following a correct response, and a less difficult item following an incorrect response—the stradaptive test continues until the termination criterion is reached. The test is terminated when a stratum is identified at which the individual is responding at or below chance level (i.e., 20% or less correct) based on a minimum of five items administered at that stratum. The individual shown in Figure 6 answered five items at stratum 8 and none of them were answered correctly. Consequently the test was terminated since further testing was likely to provide little additional information on the testee's ability level.

Scoring of the stradaptive test results in both ability level scores and consistency scores. Ability level scores reflect the individual's position on the ability scale; consistency scores reflect the variation in item difficulties encountered as the individual goes through the stradaptive test. Figure 7 shows the stradaptive test response record for an inconsistent individual. This person started the test with a relatively difficult item at stratum 8 but answered some easy items incorrectly (e.g., items 8 and 26) and some difficult items correctly (e.g., items 1 and 17). The result was a response record which varied widely across six strata. A comparison of the consistency scores for Figure 7 with those of Figure 6 shows the former to be uniformly higher. Thus, the testee depicted in Figure 7 was more inconsistent in his interaction with this item pool than was the individual in Figure 6.

Our live-testing test-retest study of the stradaptive test was based on about 200 subjects. Over an average five-week period the test-retest reliability for the best method of scoring the stradaptive test was .90; the test-retest reliability for a conventional test using the number of items administered on the average in the stradaptive test (28 items) was .86. This result showed about the same difference in favor of the adaptive test as we have obtained with other adaptive testing strategies.

I had hypothesized earlier (Weiss, 1973) that consistency scores should reflect something about the dimensionality that results from an individual's interaction with an item pool. To extend this hypothesis, if an individual is responding unidimensionally his scores should be more reliable than an individual whose interaction with an item pool is multi-dimensional. In operationalizing this hypothesis, consistency scores were used as an indicator of dimensionality, and test-retest stability as an estimate of reliability. Specifically, testees were divided into five sub-groups on the basis of their time 1 consistency scores, and test-retest reliabilities were computed separately for each of the five sub-groups. The results are shown in Table 1 for consistency score 11, the standard deviation of items encountered.

As Table 1 shows, the highest test-retest stabilities were observed for the very high consistency group for all ten methods of estimating ability within the stradaptive test. The clearest pattern emerged for ability score 1. On that score, the stability for the highly consistent testees was .94, and that for the very low consistency group was .65, with stabilities for the intermediate groups decreasing with decreasing consistency. The possible utility of consistency scores as a moderator variable is that it might permit us to make more stable predictions for some groups of individuals (consistent testees) than for others (inconsistent testees). Particularly noteworthy is the test-retest reliability of .98 for the very highly consistent testees on ability scores 8 and 9.

If these results can be replicated over longer periods of time, the consistency score might prove to be a very useful and powerful moderator variable derivable from a stradaptive testing response record. It appears to be powerful because it also moderates the test-retest reliability, but not

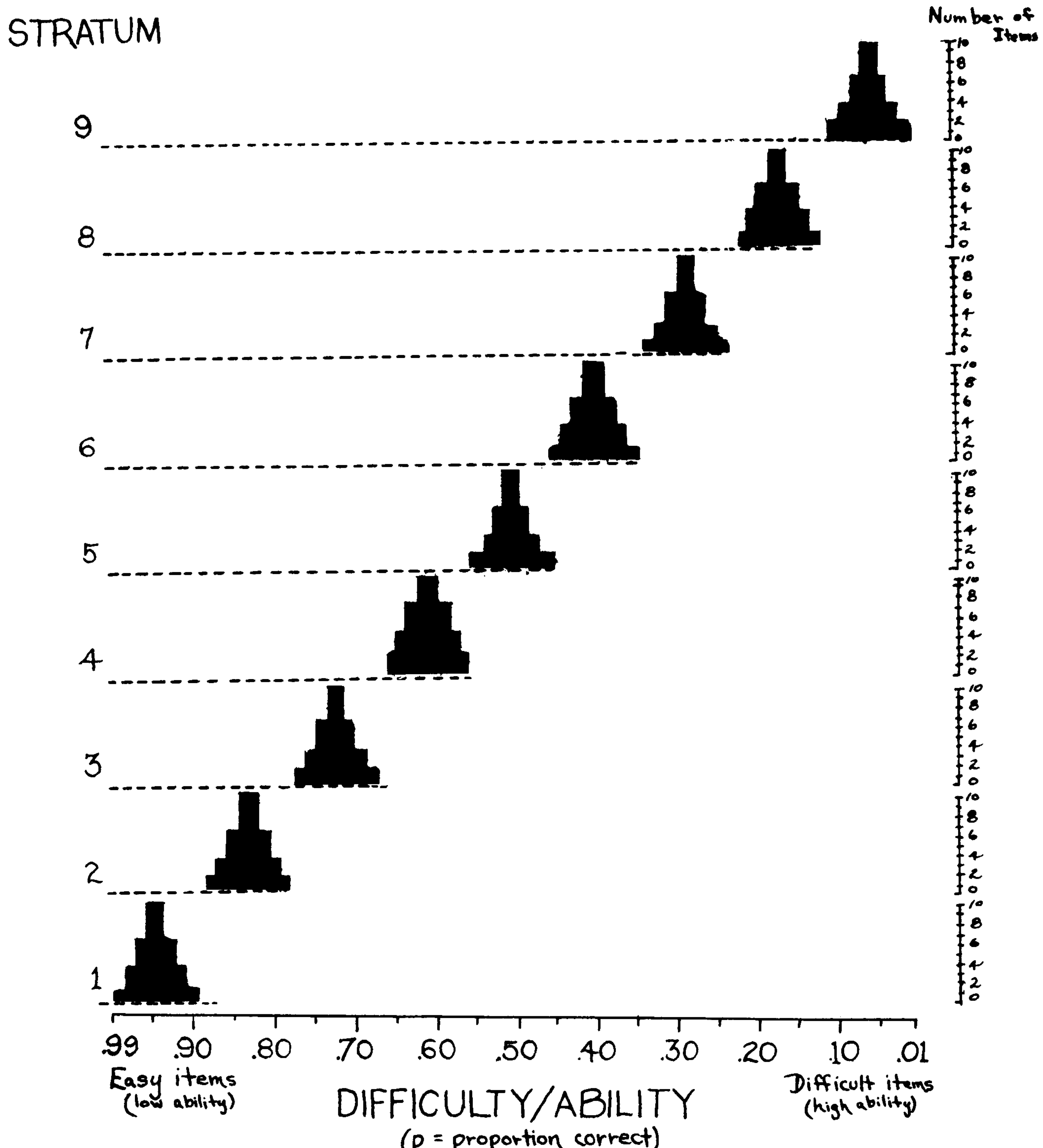
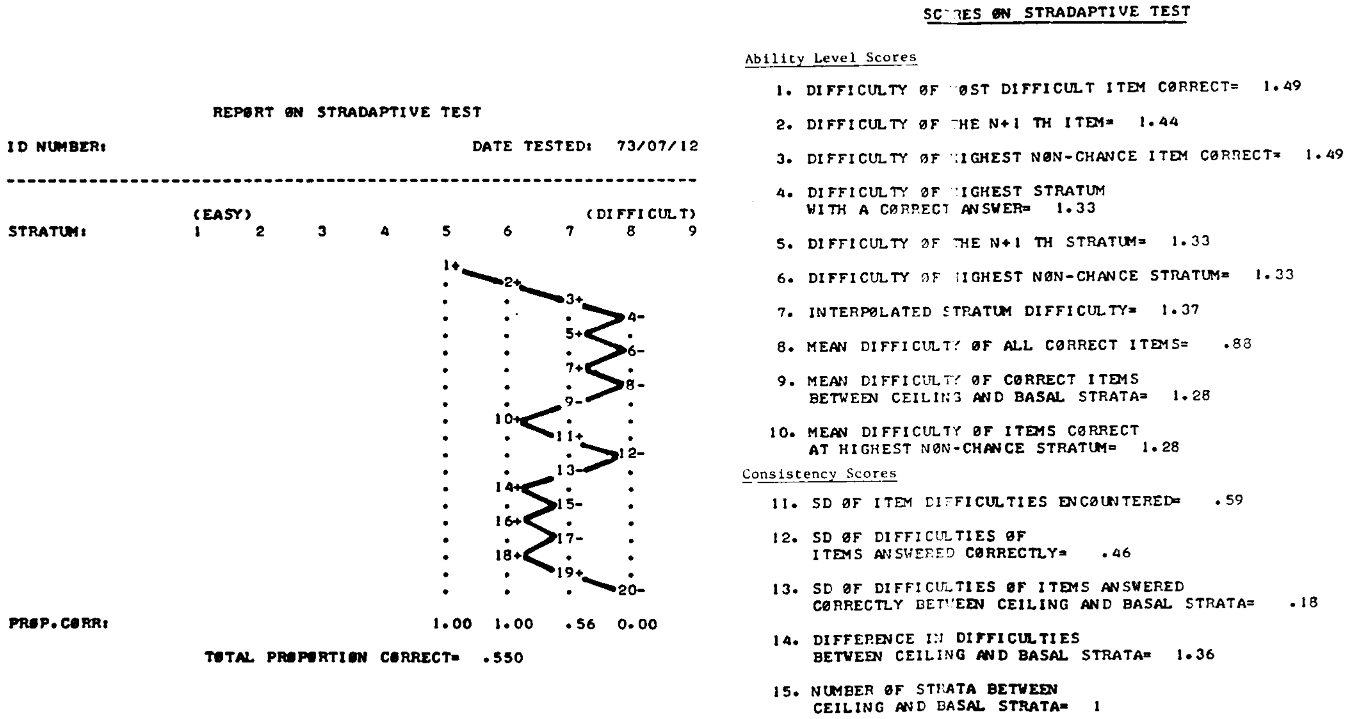


Figure 5

Distribution of Items, by Difficulty Level, in a Stradaptive Test

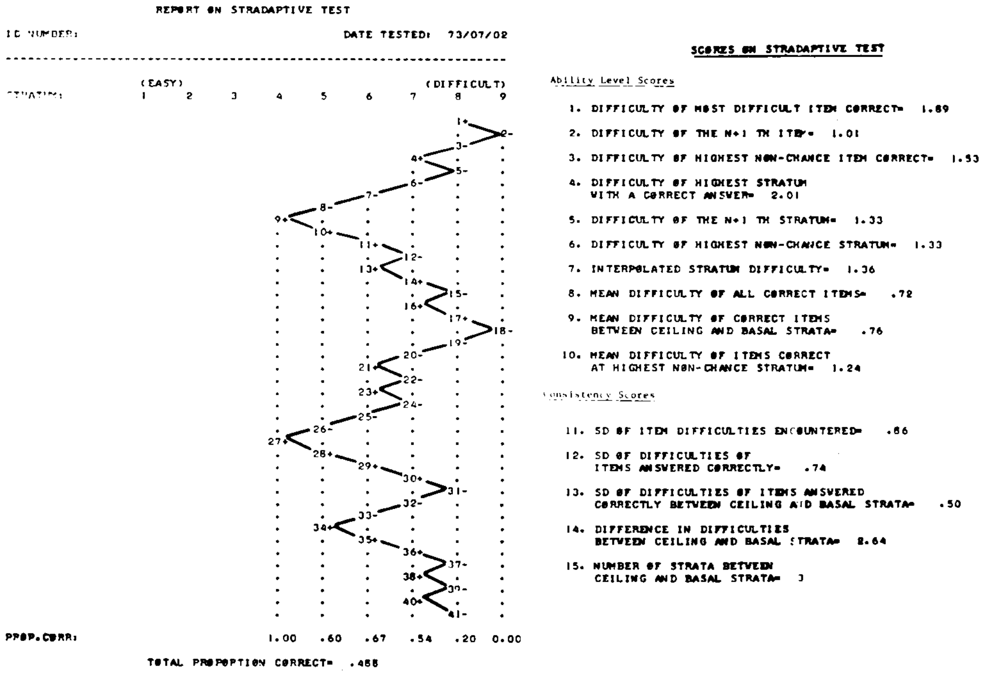


```
                         REPØRT ØN STRADAPTIVE TEST

ID NUMBER:                                         DATE TESTED:  73/07/12

-------------------------------------------------------------------------

                      (EASY)                                  (DIFFICULT)
STRATUM:               1     2     3     4     5     6     7     8     9

PRØP.CØRR:                                   1.00  1.00   .56  0.00

                TØTAL PRØPØRTIØN CØRRECT=  .550
```

```
                         SCORES ØN STRADAPTIVE TEST
Ability Level Scores
   1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  1.49
   2. DIFFICULTY ØF THE N+1 TH ITEM=  1.44
   3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49
   4. DIFFICULTY ØF HIGHEST STRATUM
      WITH A CØRRECT ANSWER=  1.33
   5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33
   6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33
   7. INTERPØLATED STRATUM DIFFICULTY=  1.37
   8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .88
   9. MEAN DIFFICULTY ØF CØRRECT ITEMS
      BETWEEN CEILING AND BASAL STRATA=  1.28
  10. MEAN DIFFICULTY ØF ITEMS CØRRECT
      AT HIGHEST NØN-CHANCE STRATUM=  1.28
Consistency Scores
  11. SD ØF ITEM DIFFICULTIES ENCØUNTERED=   .59
  12. SD ØF DIFFICULTIES ØF
      ITEMS ANSWERED CØRRECTLY=   .46
  13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED
      CØRRECTLY BETWEEN CEILING AND BASAL STRATA=   .18
  14. DIFFERENCE IN DIFFICULTIES
      BETWEEN CEILING AND BASAL STRATA=  1.36
  15. NUMBER ØF STRATA BETWEEN
      CEILING AND BASAL STRATA=  1
```

Figure 6

Report on a Stradaptive Test for a Consistent Testee



```
                         REPØRT ØN STRADAPTIVE TEST

ID NUMBER:                                         DATE TESTED:  73/07/02

-------------------------------------------------------------------------

                      (EASY)                                  (DIFFICULT)
STRATUM:               1     2     3     4     5     6     7     8     9

PRØP.CØRR:                                   1.00   .60   .67   .54   .20  0.00

                TØTAL PRØPØRTIØN CØRRECT=  .488
```

```
                         SCØRES ØN STRADAPTIVE TEST
Ability Level Scores
   1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  1.89
   2. DIFFICULTY ØF THE N+1 TH ITEM=  1.01
   3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.53
   4. DIFFICULTY ØF HIGHEST STRATUM
      WITH A CØRRECT ANSWER=  2.01
   5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33
   6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33
   7. INTERPØLATED STRATUM DIFFICULTY=  1.36
   8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .72
   9. MEAN DIFFICULTY ØF CØRRECT ITEMS
      BETWEEN CEILING AND BASAL STRATA=   .76
  10. MEAN DIFFICULTY ØF ITEMS CØRRECT
      AT HIGHEST NØN-CHANCE STRATUM=  1.24
Consistency Scores
  11. SD ØF ITEM DIFFICULTIES ENCØUNTERED=   .66
  12. SD ØF DIFFICULTIES ØF
      ITEMS ANSWERED CØRRECTLY=   .74
  13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED
      CØRRECTLY BETWEEN CEILING AND BASAL STRATA=   .50
  14. DIFFERENCE IN DIFFICULTIES
      BETWEEN CEILING AND BASAL STRATA=  2.64
  15. NUMBER ØF STRATA BETWEEN
      CEILING AND BASAL STRATA=  3
```

Figure 7

Report on a Stradaptive Test for an Inconsistent Testee

TABLE 1

STRADAPTIVE and Conventional test Test-Retest Correlations as a Function of Consistency Score 11 on Initial Testing

		Status on Consistency Score 11				
		Very High	High	Average	Low	Very Low
Mean Consistency Score		.517	.625	.706	.815	1.038
Number of Testees in Interval		27	30	41	43	29
Stradaptive Ability Score:	1	.940	.849	.847	.768	.652
	2	.875	.721	.799	.778	.751
	3	.956	.813	.878	.826	.708
	4	.934	.840	.846	.731	.664
	5	.896	.722	.793	.756	.741
	6	.950	.798	.886	.820	.704
	7	.970	.844	.902	.851	.758
	8	.981	.927	.915	.853	.869
	9	.983	.939	.907	.899	.889
	10	.951	.792	.882	.822	.718
Conventional Test		.979	.890	.918	.826	.878

as systematically, on the conventional test administered at the same time. Table 1 shows a test-retest reliability of .979 on the conventional test for the highly consistent group using the consistency scores derived from the stradaptive test. But consistency scores are not derivable from a conventional test so it is necessary to implement this finding within the framework of the stradaptive testing strategy.

Figure 8 shows a number of "subject characteristic curves," which are derivable from the stradaptive test. These curves, which reflect the individual's consistency of interaction with a stradaptive test, are based on a plot of proportion correct for each individual at each stratum of the stradaptive test. For example, the plot for "William W." shows that he answered all items correctly at both stratum 5 and stratum 6, about half correct at stratum 7 and none correct at stratum 8. Since proportion correct decreases monotonically with increasing item difficulty this individual appears to be interacting with this item pool unidimensionally; William W. is a highly consistent individual. By way of contrast, the subject characteristic curve for "Carol C." does not decrease monotonically, reflecting an inconsistent individual who answers items correctly at a variety of difficulty levels.

To be useful, these subject characteristic curves must be stable across time. To investigate their stability across an average five-week retest interval we computed canonical correlations between proportions correct at initial test and at retest. The complete redundancy analysis showed that 67% of the variance in retest subject characteristic curves was predictable from initial testing. This is equivalent to a squared multiple correlation of .82 for predicting individual proportion correct at Time 2 from a best-weighted linear combination of proportions correct at Time 1. These results imply that subject characteristic curves are reasonably stable and that they may represent a stable trait of the individual. But, certainly, more research is needed.

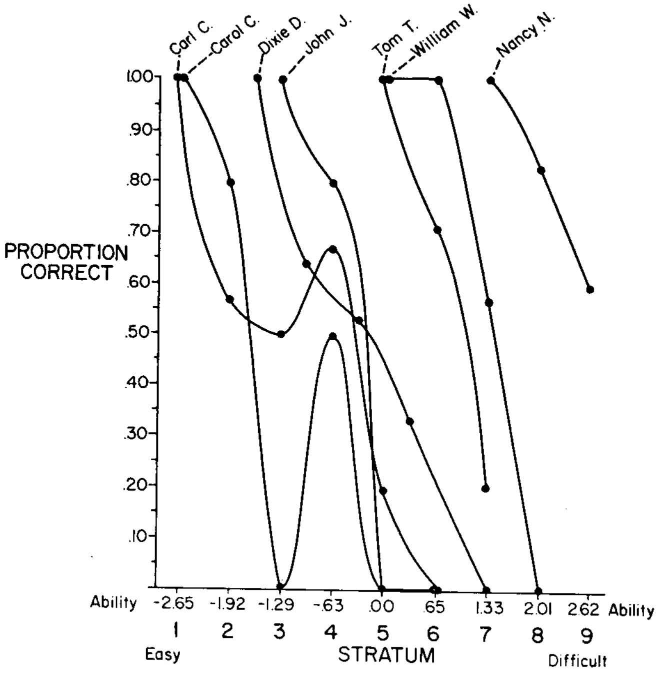


Figure 8

Proportion correct at each stratum, by individual

In addition to this live-testing study of the stradaptive test, we also have some recent data from a computer simulation study. Items with constant discriminations, and difficulties rectangularly distributed between normal ogive difficulty values of −3.33 and 3.33 and grouped into nine equally wide strata were used for the stradaptive test. Items with constant discriminations and with difficulties rectangularly distributed between −.33 and .33 (equivalent to the middle stratum of the stradaptive test) were used for the conventional test. 1000 Ss were generated with abilities in the given interval at each of 13 intervals of θ. Major findings are shown in Figure 9 and Table 2.

Figure 9 shows the information functions for the stradaptive and conventional tests at two different levels of item discrimination. At both levels of item discrimination, the information function for the stradaptive test was more horizontal than that of the conventional test, with the difference more pronounced at the higher level of item discrimination. In confirmation of Lord's theoretical predictions, the conventional test has a higher information function than the stradaptive test at the center of the ability distribution, but the range of superiority diminishes with increasing item discriminations. However, the information function for the stradaptive test increases with ability level, and for the lower discriminating items, the stradaptive test at $\theta \geqslant 2.5$ yields a higher information function than the highest value reached by the conventional test.

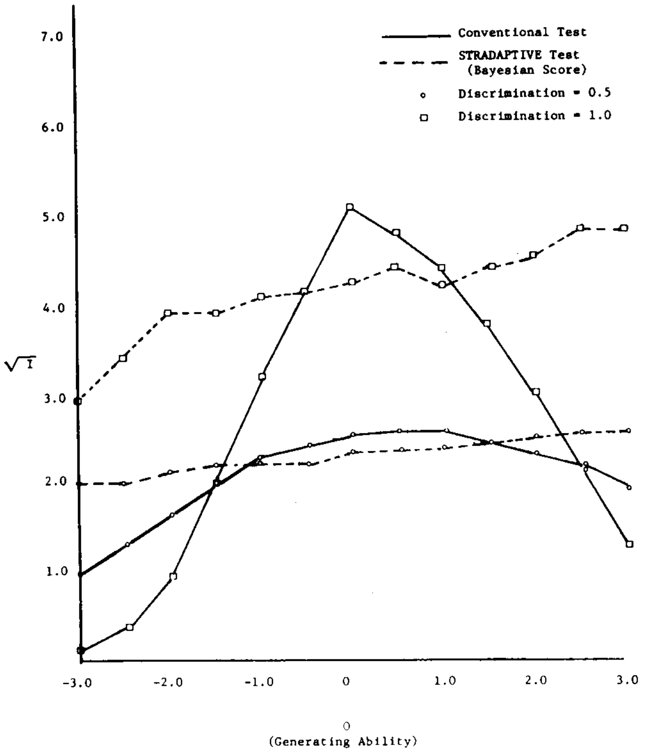


Figure 9

Information Functions for 60-Item Tests

Table 2 shows validities—correlations of ability estimate and generated ability—from the simulation data on conventional and stradaptive tests. Validity correlations are shown as a function of both item discriminations and number of items. These results show a slight superiority in validities for the conventinal tests when item discriminations are low (a=.5), and there are 40 or fewer items in both tests; a similar result is found for 10-item tests composed of items at a=1.0. In all other conditions, the stradaptive test yields higher validity, with sizable differences appearing as number of items increases and discriminations increase. For 60-item tests at a=2.0, the validity of the stradaptive test was r=.989, while the conventional test validity was only .926.

Thus, the data from both the live-testing study and the simulation study of stradaptive tests show that the stradaptive test yields scores which are more equi-precise across the ability range, and have higher validities and reliabilities than conventional tests under certain conditions. Further, the stradaptive test consistency scores appear to be powerful moderator variables which may have important practical applications in testing individuals.

Psychological effects of computerized administration. One of the psychological variables that has been unsystematically manipulated in computerized testing studies has been feedback or knowledge of results. In computerized testing we now have the capability to tell an individual whether his answer was correct or incorrect after each item in a test. But it is possible that such immediate knowledge of results might have an effect on test scores. Thus, we designed a pilot study to systematically manipulate feedback and study its effects on test scores.

TABLE 2

Score–Ability Correlations of the Stradaptive Bayesian Score and the Conventional Test Score for Tests of 10 to 60 Items, as a Function of Item Discrimination

	Discrimination (a)		
No. Items	0.5	1.0	2.0
10			
Strat	.689	.840	.919
Conv	.703	.851	.888
20			
Strat	.798	.918	.963
Conv	.811	.908	.906
40			
Strat	.869	.955	.983
Conv	.887	.938	.918
60			
Strat	.920	.971	.989
Conv	.917	.950	.926

We administered two tests on the computer to a group of inner-city high school students. The group was racially mixed, consisting of both white students and black students. Both a conventional test and a pyramidal adaptive test were administered to each student, and half the group received the conventional test first and half received the adaptive test first. In addition, half the group received feedback after each item and the other half received no feedback after each test item. We analyzed the data for the conventional test only—thus, the dependent variable in this analysis was number correct on the conventional test. The design was a 2x2x2 analysis of variance. The independent variables were 1) race—black and white; 2) feedback—immediate or none; and 3) order—conventional test administered first or second in the pair.

In order to make the feedback relevant to the high school group, we had previously asked a subgroup of students from the same school to generate a set of statements which would, to them, indicate that they answered an item correctly. We used six such statements, in pseudorandom order, including "right on," "that's cool, now try this one." and "all right, how about this one." This was done on the hypothesis that feedback can have an effect only if it is meaningful or relevant to the testee.

The results for the three-way analysis of variance are shown in Table 3. The only significant main effect was for race. Mean scores for the blacks was 17.74 and that for the whites was 27.92, on the 40-item test. Neither order nor

TABLE 3

Mean Test Scores for Blacks and Whites on the 40-item Test in Two Orders and With and Without Feedback

Group	Feedback		No Feedback		Total Group	
	N	Mean	N	Mean	N	Mean
Blacks–First	8	26.38	6	13.83	14	21.00
Second	7	13.86	6	14.67	13	14.23
Whites–First	15	26.07	14	30.93	29	28.41
Second	15	30.00	19	25.53	34	27.50
Blacks	15	20.53	12	14.25	27	17.74
Whites	30	28.03	33	27.82	63	27.92
First	23	26.17	20	25.80	43	26.00
Second	22	24.86	25	22.92	47	23.83
Total	45	25.53	45	24.20	90	24.87

3–Way Anova

Source of Variation	DF	Mean Square	F	Est. P
Order	1	105.76	1.36	.25
Race	1	2,013.26	25.84	<.00
Feedback	1	81.74	1.05	.31
Race x Order	1	161.54	2.07	.15
Order x Feedback	1	28.74	.37	.55
Race x Feedback	1	170.40	2.19	.14
Order x Race x Feedback	1	599.46	7.69	<.01
Error	82	77.92		

feedback effects were significant, nor were any of the two-way interactions. The three-way order x race x feedback interaction was significant at $p<.01$.

Figure 10 shows the means for the three-way interaction. As is indicated in Figure 10, under conditions of immediate feedback, when a conventional test was administered first, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0) who completed the conventional test under the same set of conditions. This result implies, if it can be replicated, that race differences observed in test scores may be a function not of differences in ability but of differences in the psychological effects of the conditions of administration. Although these findings do not completely replicate those of Johnson & Mihal (1973), they do support their general conclusion that conditions of test administration might affect motivational conditions, which in turn reduce race group differences to nonsignificant levels.

There is some data in our results which suggest that the three-way interaction results might be due to motivational effects. In addition to analyzing test scores, we also anlayzed the proportion of items skipped on the conventional test under the two experimental conditions and for the two racial groups. These results showed that blacks skipped more items than whites, in general, but when the conventional test was administered first to the black students and they received feedback, they skipped almost no items. This is also the same set of conditions under which the test scores for the blacks were not significantly different than those of the whites. This appears to be a motivational effect since when the blacks are given feedback the test becomes relevant to them; and when it becomes relevant they can answer the questions just as well as the whites.

Future Plans

Based on these preliminary findings we plan to continue to investigate the nature of feedback effects, and the effects of other psychological variables, on test scores. We also plan to continue to study various branching schemes in an attempt to develop optimal branching schemes which result in maximum reduction in psychometric error at all ability levels. Our general goal, as I indicated earlier, is to explore all aspects of computerized ability testing in an effort to make maximal use of the computer as a vehicle for making each individual's test score as error-free as possible.

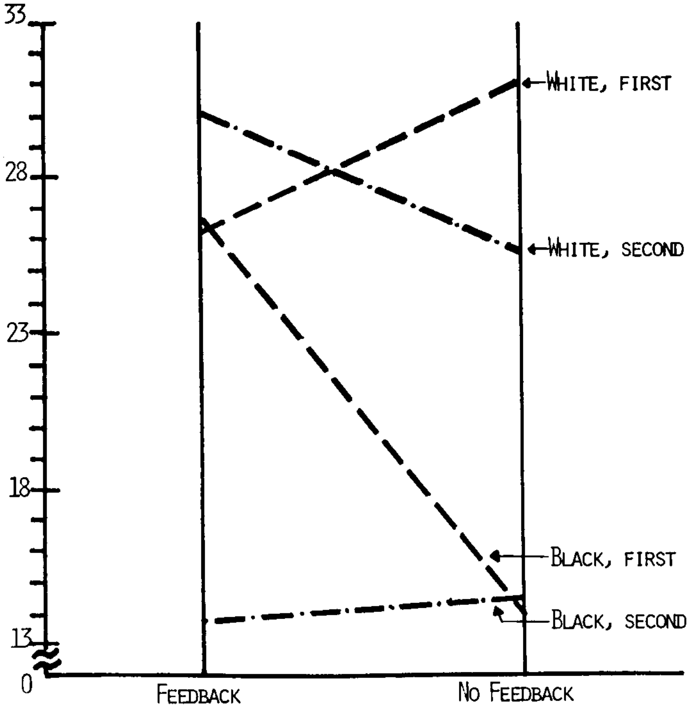


Figure 10

Mean Scores for Blacks and Whites
Completing the 40-item Test First and
Second in Both Feedback Conditions

REFERENCES

Betz, N. E. & Weiss, D. J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, October 1973.

Betz, N. E. & Weiss, D. J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Min.iesota, October 1974.

Betz, N. E. & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3, Psychometric Methods Program Department of Psychology, University of Minnesota, July 1975.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766.

Johnson, D. I. & Mihal, W. M. Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist,* 1973, *28,* 694-699.

Larkin, K. C. & Weiss, D. J. An empirical investigation of computer-administered pyramidal testing. Research Report 74-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, July 1974.

Larkin,K. C. & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1975.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement,* 1971, *8* 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement,* 1971, *31,* 805-813. (b)

Lord, F. M. Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association,* 1971, *66,* 707-711. (c)

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, in press.

Urry, V. W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: Bureau of Testing, University of Washinton, 1971.

Weiss, D. J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, September 1973.

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5, Psychometric Methods Program, University of Minnesota, December 1974.

Weiss, D. J. & Betz, N. E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1973.

ADAPTIVE TESTING RESEARCH AT MINNESOTA – SOME PROPERTIES OF A BAYESIAN SEQUENTIAL ADAPTIVE MENTAL TESTING STRATEGY[1]

JAMES R. MCBRIDE
University of Minnesota

Adaptive or tailored testing subsumes a number of different strategies for adapting the difficulty of test items to the ability of the examinee. One of the most elegant of such strategies is a Bayesian sequential technique proposed by Owen (1969) and studied empirically by several investigators including Wood (1969), Urry (1971) and Jensema (1972).

Owen's technique is a general one for the sequential design and the analysis of independent experiments with a dichotomous response. Its application in mental testing is to the problem of estimating ability by means of sequential selection, administration and scoring of dichotomous test items. The mathematical details of the method arise out of latent trait theory, with the item characteristic curves all assumed to take the form of the normal ogive. The properties of the normal ogive item characteristic function, and its logistic approximation, have been described by Lord & Novick (1968) and Birnbaum (1968), respectively.

Owen's procedure involves the individually tailored sequential design of a test by appropriate choice of available item parameters[2] (a_g, b_g, c_g) and estimation of ability via a Bayesian-motivated approximation. At each step m in the ability estimation sequence, a normal prior distribution on ability (θ) is assumed, with parameters ($\mu_m, \sigma^2{}_m$), where m indicates the number of items already administered in the sequence. A test item to be administered at step $m+1$ is selected so as to minimize a quadratic loss function on θ. With c_g=0 (i.e., no guessing) and discrimination parameters a_g constant over items, the appropriate item is the available one which minimizes the absolute value of the difference ($b_g - \mu_m$). With $c_g > 0$ the optimal difference is somewhat negative, that is, optimal difficulty is somewhat "easier" than examinee's ability. Following item administration at step $m+1$, the parameters μ_m, $\sigma^2{}_m$ of the prior distribution are updated in accord with the examinee's performance on the item. In the case of a correct answer:

$$\mu_{m+1} = E(\theta|1) = \mu_m + (1-c_g)\left(\frac{\sigma^2{}_m}{\sqrt{\frac{1}{a_g^2} + \sigma^2{}_m}}\right)\left(\frac{\phi(D)}{c_g + (1-c_g)\,\Phi(-D)}\right) \tag{1}$$

and

$$\sigma^2{}_{m+1} = \mathrm{var}(\theta|1) = \sigma^2{}_m \left\{ 1 - \left(\frac{1-c_g}{1 + \frac{1}{a^2{}_g\sigma^2{}_m}}\right)\left(\frac{\Phi(D)}{A}\right)\left(\frac{(1-c_g)\phi(D)}{A-D}\right)\right\}$$

Following a wrong answer

$$\mu_{m+1} = E(\theta|0) = \mu_m - \left(\frac{\sigma^2{}_m}{\sqrt{\frac{1}{a_g^2} + \sigma_m^2}}\right)\left(\frac{\phi(D)}{\Phi(D)}\right) \tag{2}$$

and

$$\sigma^2{}_{m+1} = \mathrm{var}(\theta|0) = \sigma^2{}_m \left\{ 1 - \left(\frac{\phi(D)}{1 + \frac{1}{a^2{}_g\sigma^2{}_m}}\right)\left(\frac{\frac{\phi(D)}{\Phi(D)} + D}{\Phi(D)}\right)\right\}$$

In the above equations (taken from Owen, 1975)

$\phi(D)$ is the normal probability density function

$\Phi(D)$ is the cumulative normal distribution function, and

$$D = (b_g - \mu_m) \Big/ \sqrt{\frac{1}{a_g^2} + \sigma_g^2} \tag{3}$$

$$A = c_g + (1-c_g)\,\phi(-D)$$

[1] Research reported herein was supported by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under contract No. 00014-67-A-0113-0029, NR No. 150-343.

Portions of these results were presented at the Spring meeting of the Psychometric Society in Iowa City, Iowa, April 1975.

A complete report of these results is in preparation (McBride & Weiss, 1975a).

[2] As most commonly used, a_g and b_g respectively are the discrimination and difficulty parameters of the normal ogive model. C_g is the guessing parameter, the probability that an examinee will respond correctly to the item when he does not know the answer. The subscript g indexes items.

μ_{m+1} and $\sigma^2{}_{m+1}$, the parameters of the Bayes posterior distribution on θ are used as the parameters of the next step's prior. At each step the prior distribution is taken to be normal, an assumption which is not strictly correct after the first item (Owen, 1975). Testing may be terminated when $\sigma^2{}_m$ becomes arbitrarily small or when m becomes arbitrarily large, or when some other criterion has been reached. At termination the latest μ_m is the estimator of θ, and $\sigma^2{}_m$ is a measure of the uncertainty of the estimate. Urry (1971) and Jensema (1972, 1974) have interpreted $\sigma^2{}_m$ as the squared standard error of estimate (S.E.E.) of θ_i. Owen (1975) gives a theorem showing that as $m \rightarrow \infty$, $\mu_m \rightarrow \theta$.

Practically speaking, of course, the number of items administered will never approach infinity; but if the pool of available items is sufficiently large and appropriately constituted, $\sigma^2{}_m$ will diminish rapidly, permitting valid estimation of θ in a very small number of items. Urry (1971, 1974) has specified the requirements for a satisfactory item pool for implementing Owen's testing procedure and has shown in computer simulation studies that Owen's sequential test can achieve in from 3 to 30 items the validity of a much longer conventional test, with the average number of items diminishing as their discriminatory power increased.

Validity, i.e., the correlation of test scores with the simulated underlying ability, is only one criterion by which to evaluate a proposed adaptive testing strategy. Since the Bayesian sequential test scores are actually estimates, in the same metric, of underlying trait level, the accuracy of the estimates is also an interesting datum. By "accuracy" here is meant the closeness of the estimates to actual ability, which may vary systematically with ability level. Another interesting property of estimates is bias, or error of central tendency. Two kinds of bias should be of some concern: 1) unconditional bias, or group mean error of estimate; and 2) conditional bias, or mean error of estimate at a given level of the parameter being estimated. As a matter of convention, then, in the following the term "accuracy" will refer to mean absolute error of estimate, $(1/N)\ \Sigma' |\hat{\theta}_i - \theta_i$; "bias" will refer to mean algebraic error of estimate $(1/N)\ \Sigma\ (\hat{\theta}_i - \theta_i)$; and "conditional bias" will refer to mean algebraic error of estimate at a given value of θ, $(1/N)\ \Sigma(\hat{\theta}_i - \theta | \theta)$.

The purpose of the present paper is to report the results of a series of simulation studies designed to investigate the influence of item pool characteristics on some properties of the Bayesian sequential test other than the correlational validity of the trait estimates. These properties will include bias and accuracy of the estimates, as well as others enumerated below.

The studies reported below were motivated by results obtained with live testing of Owen's strategy. Using a 329-item pool of vocabulary knowledge test items, a correlation of .80 was obtained between estimated ability and number of test items to termination (McBride & Weiss, 1975b). Simulation studies designed to investigate the influence of the item pool on that unexpectedly large correlation led to our discovery of systematic non-linear bias in the Bayesian estimates of ability. The nature of the bias, and some of its correlates, are discussed below.

METHOD

1. *Dependent variables* of interest included test length (number of test items administered before the termination criterion was reached), errors of estimate $(\hat{\theta} - \theta)$, bias of estimate (mean over individuals of $(\hat{\theta} - \theta)$), absolute value of the error $|\hat{\theta} - \theta|$, and validity of the estimates of θ, $r_{\theta\hat{\theta}}$.

2. *Independent variables* of interest included the effects of guessing in both the response model and the scoring algorithm, of item discrimination, and the correlation of difficulty and discrimination parameters in the item pool, and of different termination criteria.

3. *Examinees* for the first study were simulated by computer-generation of pseudorandom numbers (from a normal population with mean 0 and variance 1) which represented the ability θ_i of each examinee, i. For the second study, 100 examinees were simulated at each of 31 points on the ability continuum.

4. *Item responses* were simulated by comparing $P'_g(\theta_i)$ for each item g and examinee i with a random number e_{gi} from a rectangular distribution in the interval [0,1]. A score of 1 for examinee i on item g was assigned if $P'_g(\theta_i) \geqslant e_{gi}$. Otherwise a score of 0 was assigned.

5. *Item pools* were simulated under two different conditions:

a. A *perfect item pool* with items of constant discrimination a_g and guessing parameter c_g was simulated. Under this condition, the computer program computed the optimal difficulty b_{m+1} of the next item to administer, and a simulated item with that difficulty value was made available. This is referred to as a "perfect" item pool because in effect we have simulated an item pool in which an unlimited number of items is available at any point on the difficulty continuum. The estimated optimal difficulty of an item to administer at stage $m+1$ is equal to the current ability estimate, $\hat{\theta}_m$, when guessing is not a factor (i.e., when c_g=0). When guessing is a factor (c_g>0), the estimated optimal difficulty b_g is smaller than $\hat{\theta}_m$ by an amount which is a joint function of a_g and c_g. That is, when c_g>0 $(b_g - \hat{\theta}_m)$<0. (Actually, the true optimal difficulty is a function of a_g, c_g and the unknown parameter θ. The Bayesian sequential test procedure only estimates θ and hence estimates the optimal item difficulty. At any rate, the simulated "perfect" item pool makes available at every step m an item whose difficulty is exactly equal to the estimated optimal item difficulty based on a_g, c_g, and the then current estimate of θ).

b. A *differentially discriminating "perfect" item pool* was simulated by having unlimited item difficulties b_g available (as in a. above), but varying item discrimination systematically so that the mean a_g could be specified and the regression of a_g of item difficulty b_g could be varied. In this way it was possible to simulate item pools in which more highly discriminating items were available in some regions of the ability continuum than in others. The details of this procedure are described in Study 2, below.

6. The Bayesian sequential test was simulated by a computer program. Input variables were θ_i; the parameters μ_0 and $\sigma^2{}_0$ of the initial prior distribution on θ; the number of items to be administered to any examinee; the constant discrimination parameter a_g of the *perfect item pool* (or the mean discrimination parameter of the *differentially discriminating perfect item pool)*, along with two guessing specifications. The first, c_i, specified the propensity of the examinees to guess while the second, c_g, specified whether guessing was to be accounted for in scoring.

Study 1: The effects of guessing

For this study the "perfect" item pool was used, with two values of c_g:$c_g= \left\{ \begin{matrix} 0 \\ .20 \end{matrix} \right.$, paired with two values of the personal guessing tendency $c_i= \left\{ \begin{matrix} 0 \\ .20 \end{matrix} \right.$. Of the four possible pairwise combinations, only three were used; resulting in three sets of conditions

	c_i	c_g
no guessing	0	0
uncorrected guessing	.20	0
corrected guessing	.20	.20

In the first condition, no guessing takes place (c_i=0) and no correction for guessing enters into the scoring formula (c_g=0). In the second condition c_i=.20 (every individual i has a random chance of correct response equal to .20), but c_g=0 (guessing goes uncorrected in the scoring algorithm). Finally, in the third condition, the .20 guessing parameter and the scoring correction for guessing take the same value.

In each condition, the same 100 "examinees" (θ_i sampled from a normal (0,1) population) were administered 14 simulated Bayesian sequential tests in which testing terminated for an examinee whenever the $\sigma^2{}_m$, the estimated variance of the posterior distribution of θ, fell below .0625 (this is equivalent to the Urry/Jensema criterion of SEE <.25). The 14 simulated tests in each condition were experimentally independent, and differed from each other in the value of the a_g parameter, which was constant within a test, but which varied systematically across tests. The 14 a_g values were a_g = .5, .6, .7, .8, .9, 1.0, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00.

For each test in each condition, the following variables were observed:

a. mean and range of test length, k

b. errors of estimate, $e_i = (\hat{\theta}_i - \theta_i)$

c. test bias, $(1/N) \sum_i (\hat{\theta}_i - \theta_i)$

d. mean absolute error, $(1/N) \sum_i |\hat{\theta}_i - \theta_i|$

e. test validity $r_{\theta\hat{\theta}}$

f. correlated error $r_{\hat{\theta}e}$ and $r_{\theta e}$

g. correlated test length $r_{\theta k}$ and $r_{\hat{\theta}k}$

Study 2: *The effects of the configuration of item parameters in the item pool*

Most simulation studies of Owen's sequential test have used a constant item discrimination parameter within each test. Typical item pools in actual use, however, have varying item discriminations, with the potential effect of having more discriminating items available in some ranges of the trait level than in others. In this study, different item pool a_gx b_g configurations were simulated by using the differentially discriminating "perfect" item pool. The approximate correlation (r_{ab}) between item discriminating power and item difficulty was varied in order to observe its effect on some properties of the Bayesian test and of the resulting scores.

Three different values of r_{ab} were simulated: -.71, 0 and +.71. With r_{ab}=.71, more discriminating items are available, on the average, at higher levels of θ.With r_{ab}=-.71 the more discriminating items were available at the lower levels of θ. And with $r_{ab} \doteq 0$, no level of θ was favored in terms of available discriminating power of the items, although discriminating power was free to vary randomly. In each "item pool" configuration, the mean item discrimination $\bar{a}_g$ was set at 1.25. Additionally, a minimum a_g value of .80 was imposed, in accord with Urry's (1974) recommendation.

The item pool configuration was simulated by means of:

1) selecting the appropriate b_g for the next item from the perfect item pool as though all a_g were equal to $\bar{a}_g$; call this $b^*{}_g = (b_g | \hat{\theta}_m, \bar{a}_g)$;

2) calculating a conditional a_g value from a linear transform of $b^*{}_g$:

$$a_g | b^*{}_g \doteq r_{ab} \left(\frac{\text{S.D.}_\text{A}}{\text{S.D.}_\text{B}}\right). b^*{}_g + \bar{a}_g$$

where S.D.$_\text{A}$ is the standard deviation of the a_g parameters in the simulated pool

S.D.$_\text{B}$ is the standard deviation of the b_g parameters in the simulated pool

a_b, $b^*{}_g$, r_{ab}, $\bar{a}_g$ are as previously defined;

3) adding an error component, e_g, to the approximate a_g, so that for each item administered $a^*{}_g = a_g | b^*{}_g + e_g$

where $a^*{}_g$ is the simulated discriminating power of the item

$a_g | b^*{}_g$ is the approximate discrimination defined above

eg is a random number from a population normal in $(0,\sigma^2{}_e)$

$$\sigma_e = \sqrt{\sigma^2{}_e} = \text{S.D.}_\text{A} \, (1-r^2{}_{ab})^{1/2}.$$

4) setting $a^*{}_g$ equal to .80 whenever it would otherwise have a lower value.

"Examinees" for this study were 3100 simulated θ's, 100 at each of 31 equally spaced intervals between -3.0

and 3.0, inclusive. The corrected guessing condition (c_g=c_i=.20) was in effect. The posterior variance termination criterion ($\sigma^2{}_m \leqslant .0625$) was used, with an arbitrary 30-item maximum test length. At each of the 31 θ levels the following variables were observed for each individual, i:

a. test length, k_i
b. test score, $\hat{\theta}_i$
c. error of estimate, $e_i = \hat{\theta}_i - \theta$

While study 1 examined average characteristics of the Bayesian test and test scores, Study 2 was concerned with certain properties of the procedure as a function of trait level, θ, and of the item pool configuration, r_{ab}. For each configuration, the regressions of k, e and $\hat{\theta}$ on θ were estimated from the means of the 100 individuals at each level of θ.

Additionally, the data were used to calculate empirical values of the information function $I_{\hat{\theta}}(\theta)$ of the Bayesian test scores θ. The information at any level θ_i may be calculated as the square of the ratio of the partial derivative with respect to θ of the regression of test scores $\hat{\theta}$ on θ, to the conditional standard deviation $(\sigma_{\hat{\theta}\,|\,\theta})$ of the test scores at the given level of θ. This may be written $I_{\hat{\theta}}(\theta) = \left[\frac{\partial/\partial\theta(\mathrm{E}(\hat{\theta}\,|\,\theta))}{\sigma_{\hat{\theta}\,|\,\theta}}\right]^2$ (after Lord, 1970, p. 153). In each configuration for each of the 31 levels of θ, the conditional standard deviation was estimated as the observed S.D. of the 100 test scores at that level. The numerator of the equation was calculated for each θ point from a third degree polynomial equation for the regression of $\hat{\theta}$ on θ, estimated by least squares fit to the thirty-one mean $\hat{\theta}$'s observed under each item pool configuration.

RESULTS

Study 1

Tables 1, 2 and 3 and Figures 1, 2 and 3 contain the results of sequential testing under the three conditions of guessing/correction for guessing, at each of 14 item discrimination levels. Some noteworthy trends are:

a. *Test length* was constant at each a_g level in the no guessing (Table 1; Figure 1) and uncorrected guessing (Table 2; Figure 2) conditions, with test length to termination diminishing proportionately with the inverse of the a_g level.

In the corrected guessing condition (Table 3 and Figure 3) test length varied across individuals, while *mean* test length within a_g level behaved in the same manner as dic test length in the other two conditions. One datum of note is the behavior of test length as a function of a_g level: in order for all examinees to reach normal termination in less than 30 items (in the corrected guessing condition), the item discrimination value must exceed 1.25 (a_g>1.25).

Another result of interest is an expected one: the corrected guessing condition required more items to termination than did the other conditions.

b. *Errors of estimate*, $e_i = (\hat{\theta}_i - \theta_i)$, were moderately correlated with ability θ and test score $\hat{\theta}$ under all conditions, as revealed in Tables 1, 2 and 3. e_i tends to be positive for θ_i<0 and negative for θ_i>0. This result was consistent, and reflects a regression effect caused by the quadratic loss function employed in the item selection procedures.

c. Test bias, mean absolute error, test validity, correlated errors and correlated test length values for the no guessing, uncorrected guessing and corrected guessing conditions are listed in Table 1, 2 and 3, respectively. Additionally, Figures 1, 2 and 3 graph some of these values as a function of a_g level within each condition. Noteworthy in these data is the sizeable bias and mean absolute error in the uncorrected guessing condition (Table 2; Figure 2), as well as the tendency for bias and absolute error to increase at a_g levels above 2.00 in the corrected guessing condition (Table 3; Figure 3). Note also that in the uncorrected guessing condition (Table 2), test validity, $r_{\hat{\theta}\theta}$, decreased at a_g levels beyond 2.00. Jensema (1972) observed this phenomenon, which he termed "correlation drop-off."

Study 2

Table 4 lists the observed mean values under each item pool configuration of test score, test length, and error of estimate for each value of θ. Figures 4, 5 and 6 depict these data graphically.

a. *Test length.* Mean test length (Figure 4) did not vary with θ in the r_{ab}0 configuration since the maximum of 30 items occurred at all levels. In the r_{ab}-.71 configuration, mean test length covaried positively and almost perfectly with ability level. In the r_{ab}+.71 configuration, test length covaried inversely with trait level, with more items required at the lower trait levels until the arbitrary 30-item limit was reached.

b. *Test scores.* The regression of mean trait estimates, $\hat{\theta}$ on θ was virtually linear in all three configurations in the interval [$-1.5<\theta<2.0$]. As can be seen from Figure 5, the Bayesian test scores tended to underestimate θ at high trait levels, and to overestimate θ at low trait levels. The regression of $\hat{\theta}$ on θ departed from a linear regression at extreme levels of θ (beyond θ = ±2.00) with the departure more noticeable in the lower extremes of the scale.

c. *Errors of estimate.* The regression of mean errors of estimate on θ, seen in Figure 6, clearly illustrates a tendency of the Bayesian test scores to overestimate θ markedly and consistently at $\theta<-1.5$ in all three item pool configurations. The tendency to underestimate high θ's is also illustrated. In this data the latter tendency was quite strong with r_{ab}-.71 but less so with r_{ab}+.71.

Information. The estimated values of the derivative $\frac{\partial}{\partial\theta}[\mathrm{E}(\hat{\theta}\,|\theta)]$, the conditional standard deviation $\sigma_{\hat{\theta}\,|\theta}$ and the information at each level of θ, under each item pool configuration, are listed in Table 5. Smoothed information curves for all three configurations are plotted in Figure 7. Some noteworthy trends are pointed out here.

TABLE 1

Test Length, Mean Errors of Estimate, and Correlates of Ability θ and Test Score θ, as a Function of Item Discrimination a_g in the Perfect Item Pool. No Guessing Condition (c_g=c_i=0).

	Item Discrimination (a_g)													
Property	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
Test Length														
Mean	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Minimum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Maximum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Error of Estimate														
Mean (Bias)	.00	-.01	.02	.01	.00	.01	.00	.02	.04	.06	.04	.05	.03	.04
Mean Absolute Error	.17	.17	.19	.19	.18	.19	.18	.21	.20	.21	.21	.20	.21	.22
Correlates														
with error														
$r_{\theta e}$	-.35	-.27	-.31	-.36	-.39	-.35	-.37	-.37	-.30	-.37	-.39	-.36	-.32	-.35
$r_{\theta e}$	-.17	-.08	-.10	-.16	-.20	-.15	-.17	-.14	-.07	-.15	-.16	-.14	-.09	-.10
with test length														
$r_{\theta k}$	...[a]	...	...	...	...	...	...	...	...	...	...	...	...	...
$r_{\theta k}$	...	...	...	...	...	...	...	...	...	...	...	...	...	...
$r_{\theta\hat{\theta}}$(validity)	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97	.97	.97	.97	.97

a. Correlations not computed since test length (k) was constant.

TABLE 2

Observed Properties of the Bayesian Sequential Test as a Function of Item Discrimination in the Perfect Item Pool. Uncorrected Guessing (c_g=0; c_i=.20`

	Item Discrimination (a_g)													
Property	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
Test Length														
Mean	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Minimum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Maximum	100	71	52	41	33	27	18	13	11	9	7	7	6	5
Errors of Estimate														
Mean (Bias)	.57	.48	.47	.42	.37	.34	.30	.27	.29	.31	.32	.31	.29	.29
Mean Absolute Error	.58	.48	.48	.46	.42	.39	.37	.37	.36	.40	.39	.38	.37	.39
Correlates														
with error														
$r_{\theta e}$	-.51	-.46	-.49	-.48	-.48	-.43	-.44	-.36	-.31	-.31	-.32	-.32	-.32	-.32
$r_{\hat{\theta} e}$	-.29	-.23	-.23	-.19	-.20	-.13	-.16	-.04	.01	.05	.05	.05	.07	.02
with test length														
$r_{\theta k}$	....[a]	....	....	....	....	....	....	....	....	....	....	....	....	....
$r_{\hat{\theta} k}$	....	....	....	....	....	....	....	....	....	....	....	....	....	....
$r_{\theta\hat{\theta}}$(validity)	.97	.97	.96	.95	.95	.95	.96	.94	.95	.93	.93	.93	.92	.91

a. Correlations not computed since test length (k) was constant.

TABLE 3

Observed Properties of the Bayesian Sequential Test as a Function of Item Discrimination in the Perfect Item Pool. Corrected Guessing (c_g=c_i=.20)

	Item Discrimination (a_g)													
Property	.5	.6	.7	.8	.9	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
Test Length														
Mean	100	99	77	60	48	40	27	20	16	13	11	10	9	9
Minimum	100	93	66	52	42	33	21	14	11	8	7	6	6	5
Maximum	100	100	88	69	57	49	32	26	21	19	18	16	15	14
Errors of Estimate														
Mean (Bias)	.04	.03	.02	.03	.02	.04	.01	.01	.01	.02	.04	.06	.07	.08
Mean Absolute Error	.22	.18	.16	.18	.19	.19	.16	.17	.19	.20	.18	.20	.19	.21
Correlates														
$r_{\theta e}$	-.39	-.36	-.25	-.39	-.42	-.35	-.37	-.37	-.38	-.39	-.25	-.37	-.33	-.33
$r_{\hat{\theta} e}$	-.17	-.18	-.09	-.20	-.23	-.16	-.19	-.18	-.18	-.19	-.14	-.14	-.10	-.08
$r_{\theta k}$	....[a]	.54	.80	.78	.78	.81	.81	.82	.85	.88	.85	.88	.90	.88
$r_{\hat{\theta} k}$	....	.56	.82	.81	.80	.83	.82	.84	.87	.89	.86	.90	.91	.90
$r_{\theta\hat{\theta}}$	.97	.98	.99	.98	.98	.98	.98	.98	.98	.98	.98	.97	.97	.97

[a] Correlations not computed since test length (k) was constant.

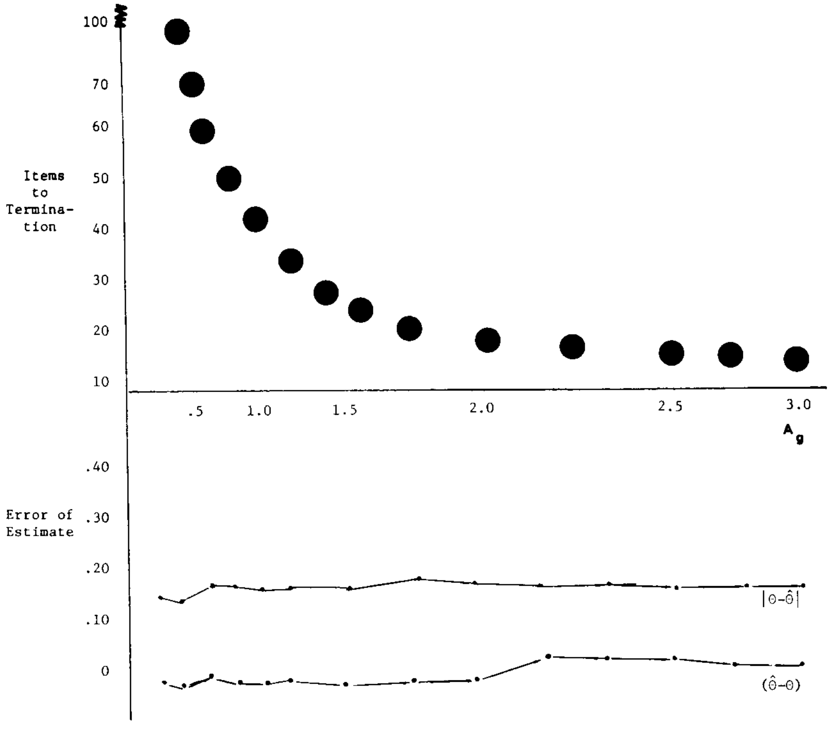


Figure 1. Some observed properties of a Bayesian sequential test, as a function of item discrimination. No guessing; perfect item pool; posterior variance termination criterion.

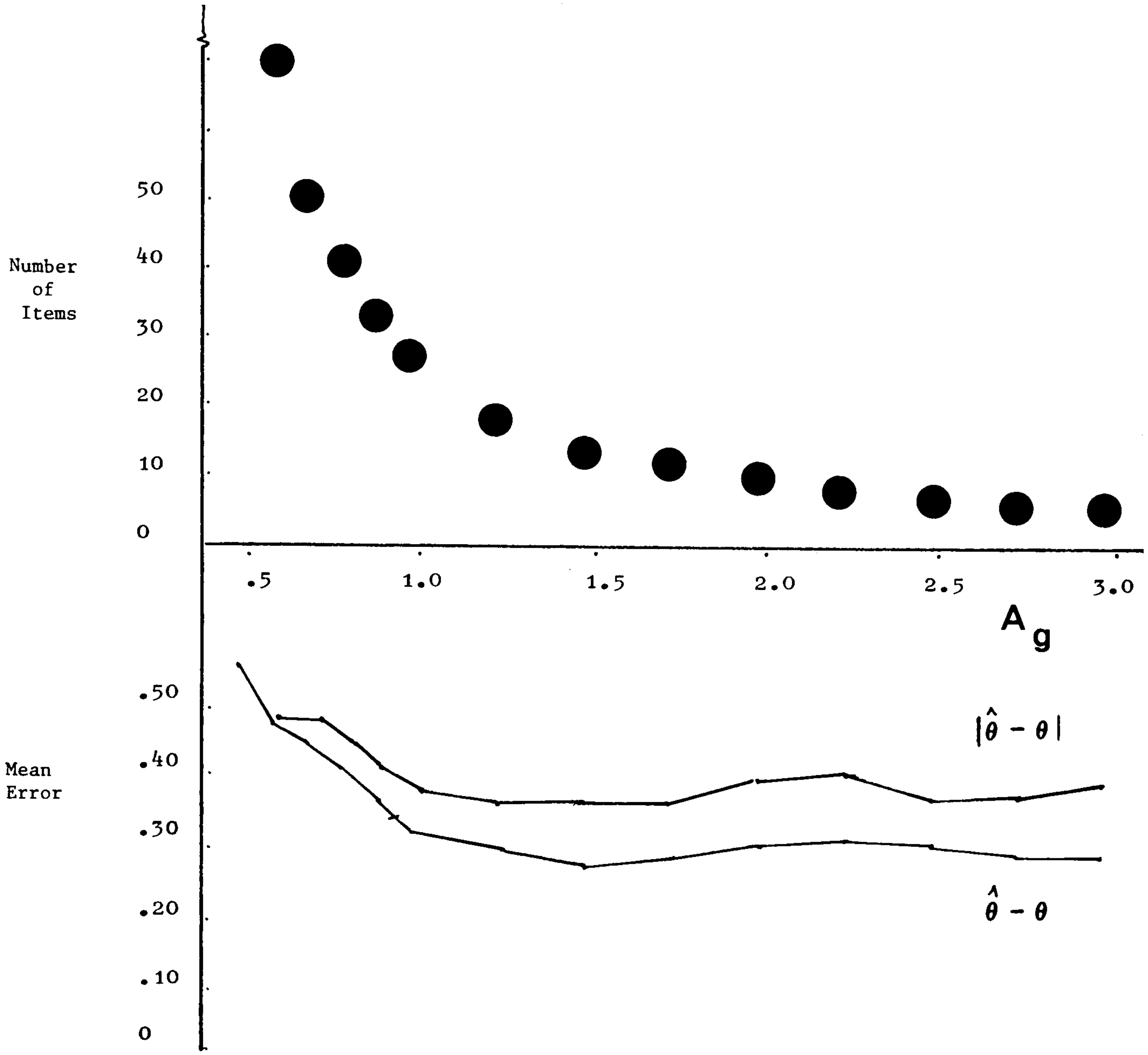


Figure 2. Some observed properties of a Bayesian sequential test, as a function of item discrimination. Uncorrected .20 guessing; perfect item pool; posterior variance termination criterion.

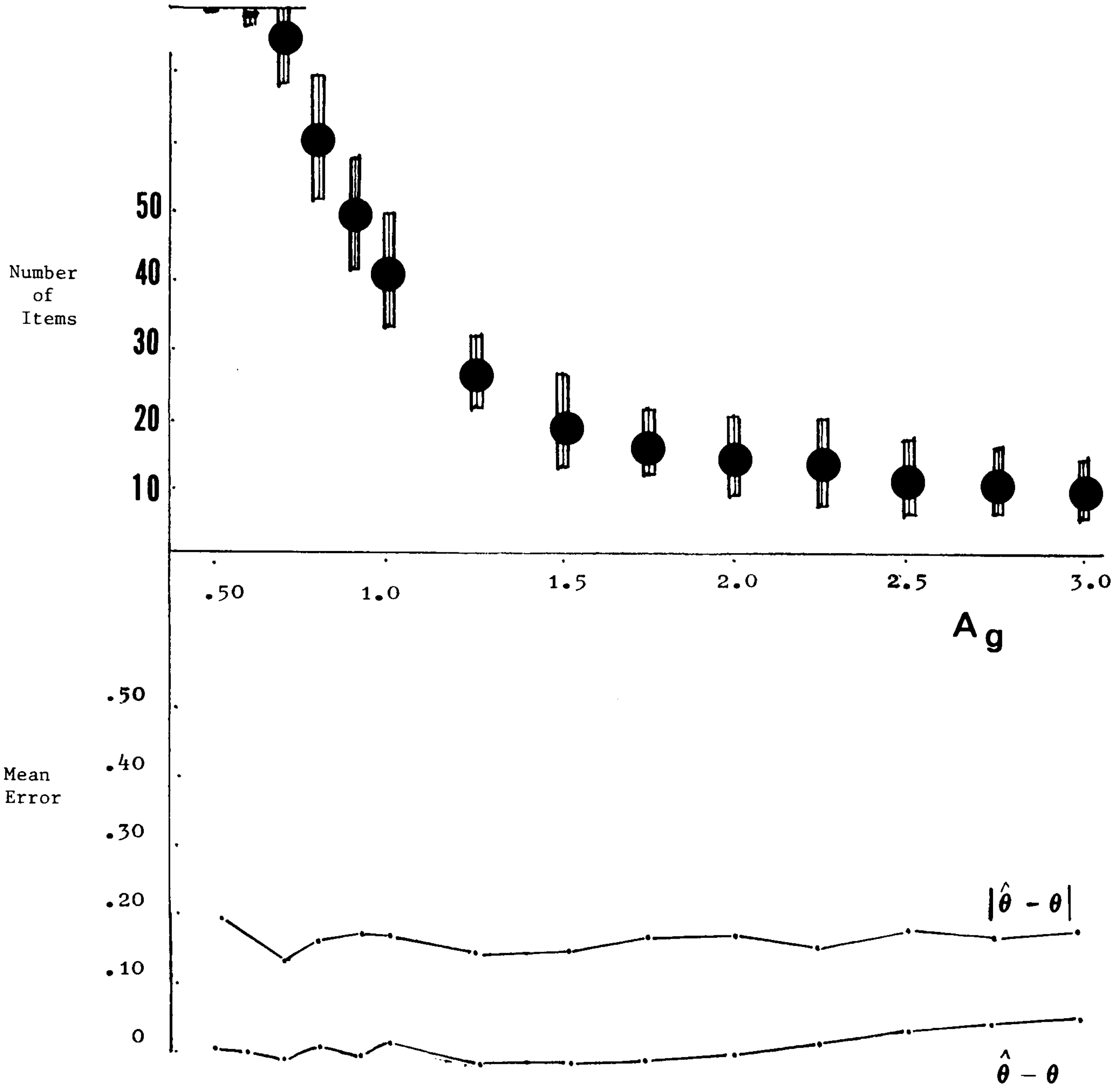


Figure 3. Some observed properties of a Bayesian sequential test, as a function of item discrimination. Corrected .20 guessing; perfect item pool; posterior variance termination criterion.

TABLE 4

Mean Test Scores ($\hat{\theta}$), Mean Test Length (k) and Mean Error of Estimate (e) for Three Item Pool Configurations, at each of 31 Trait Levels (θ)

	Item Pool Configurations								
	r_{ab}+.71			r_{ab}.0			r_{ab}-.71		
θ	$\hat{\theta}$	k	e	$\hat{\theta}$	k	e	$\hat{\theta}$	k	e
-3.0	-2.39	30	.612	-2.47	30	.532	-2.30	14	.696
-2.8	-2.26	30	.545	-2.29	30	.513	-2.20	14	.601
-2.6	-2.06	30	.542	-2.25	30	.352	-2.17	15	.427
-2.4	-2.00	30	.404	-2.06	30	.342	-2.08	15	.317
-2.2	-1.81	30	.390	-1.94	30	.263	-1.93	16	.269
-2.0	-1.70	30	.296	-1.80	30	.204	-1.74	17	.263
-1.8	-1.60	30	.200	-1.66	30	.141	-1.65	18	.146
-1.6	-1.44	30	.163	-1.45	30	.151	-1.48	18	.125
-1.4	-1.24	30	.162	-1.32	30	.082	-1.29	20	.110
-1.2	-1.12	30	.076	-1.12	30	.082	-1.14	21	.060
-1.0	- .93	30	.073	- .93	30	.071	- .98	22	.018
- .8	- .74	30	.055	- .74	30	.055	- .76	24	.037
- .6	- .56	30	.038	- .59	30	.014	- .58	26	.015
- .4	- .44	30	-.040	- .40	30	.004	- .35	27	.049
- .2	- .25	30	-.046	- .21	30	-.010	- .14	29	.062
0	- .06	30	-.058	.05	30	.046	.02	30	.021
.2	.20	30	-.003	.16	30	-.039	.19	30	-.007
.4	.35	30	-.053	.34	30	-.056	.35	30	-.051
.6	.53	29	-.068	.61	30	.010	.58	30	-.015
.8	.76	29	-.044	.74	30	-.058	.81	30	.013
1.0	.95	28	-.051	.89	30	-.113	.92	30	-.080
1.2	1.11	27	-.091	1.16	30	-.036	1.15	30	-.047
1.4	1.37	26	-.034	1.33	30	-.068	1.25	30	-.150
1.6	1.53	26	-.074	1.48	30	-.117	1.46	30	-.140
1.8	1.73	25	-.070	1.68	30	-.123	1.64	30	-.165
2.0	1.89	24	-.113	1.88	30	-.119	1.78	30	-.224
2.2	2.09	24	-.107	2.05	30	-.146	1.98	30	-.224
2.4	2.27	23	-.132	2.22	30	-.176	2.13	30	-.270
2.6	2.47	23	-.126	2.37	30	-.230	2.33	30	-.273
2.8	2.63	23	-.168	2.57	30	-.230	2.43	30	-.372
3.0	2.81	23	-.189	2.72	30	-.282	2.57	30	-.426

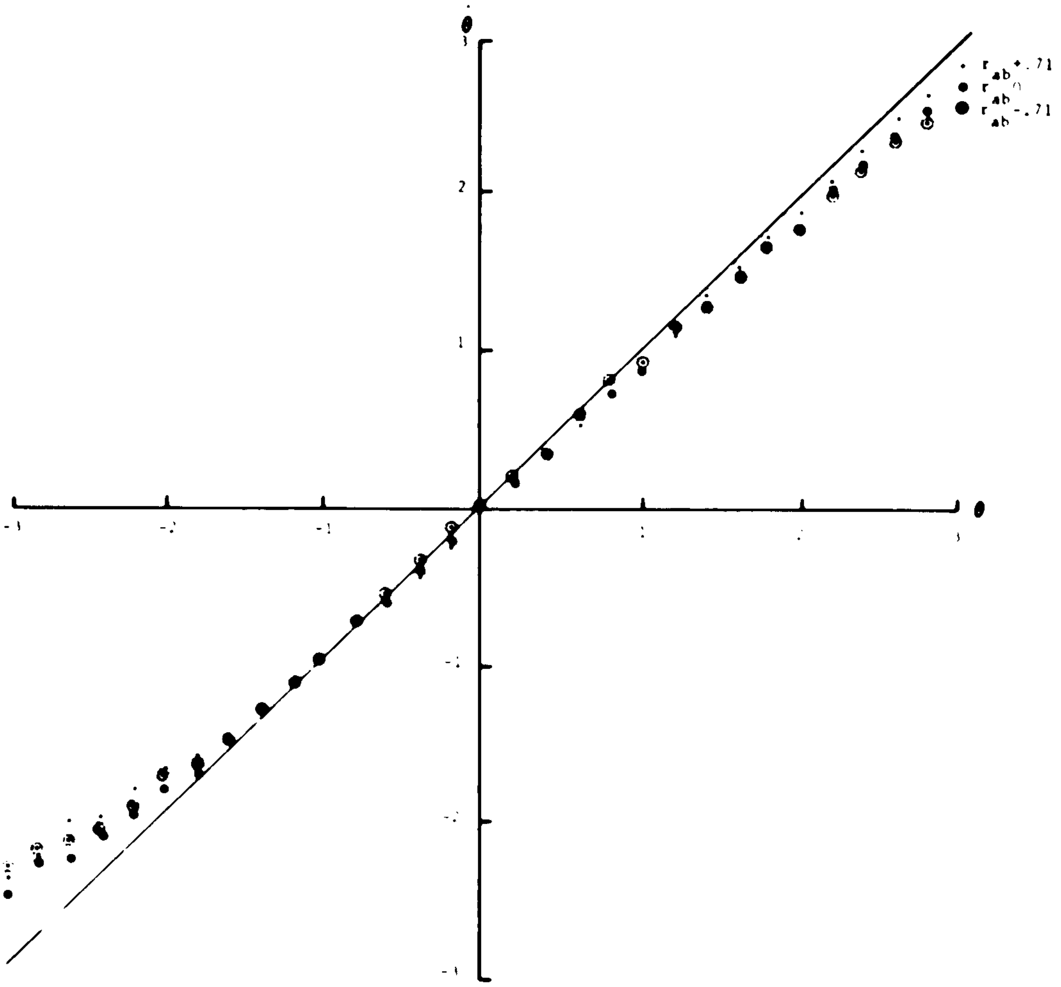


Figure 4. Mean estimated ability ($\hat{\theta}$) at thirty-one ability points (θ) for the simulated Bayesian sequential test under three item pool configurations.

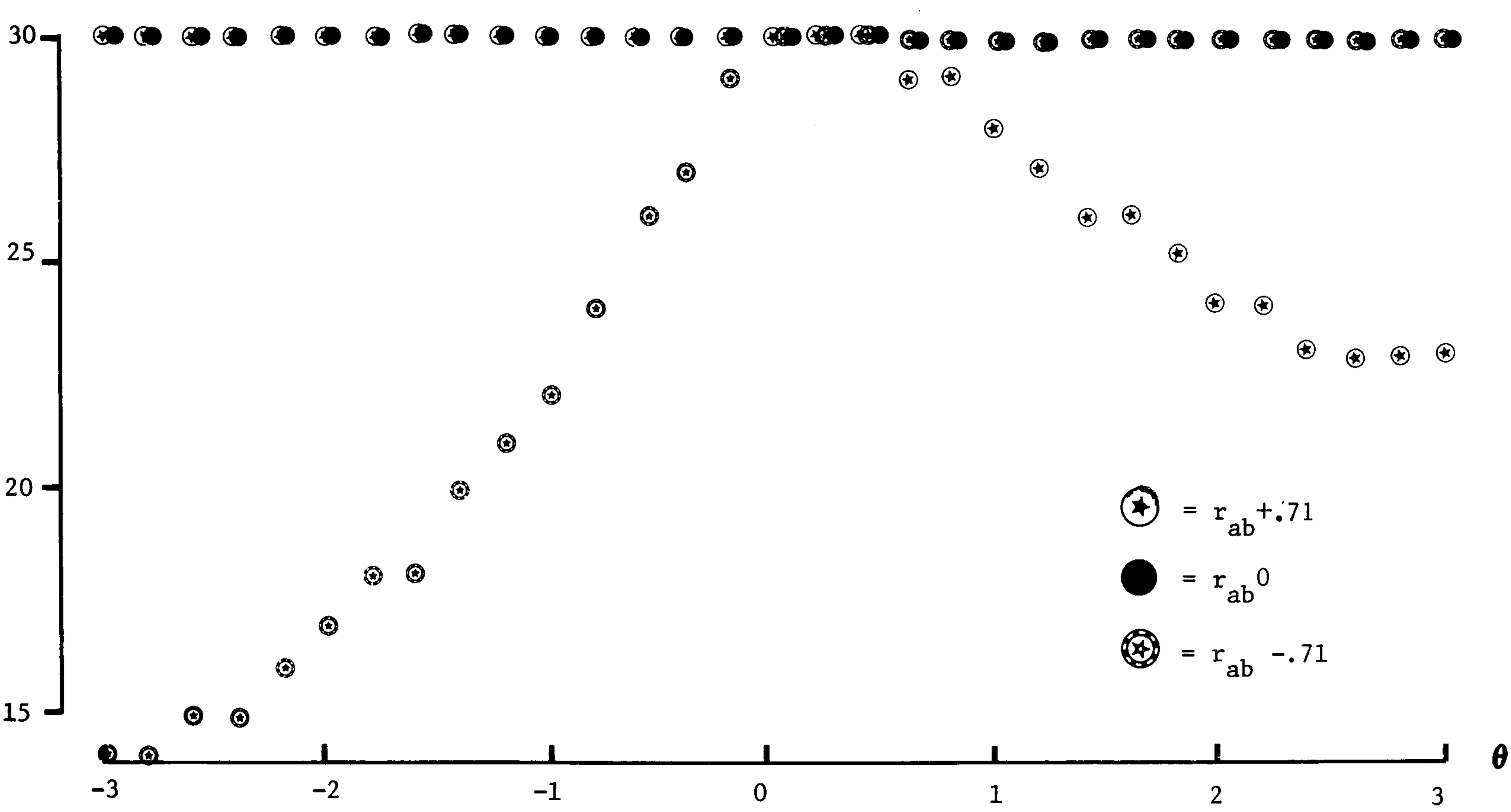


Figure 5. Mean number of items to termination (test length) at thirty-one ability points (θ) for the simulated sequential test under three item pool configurations (See text.)

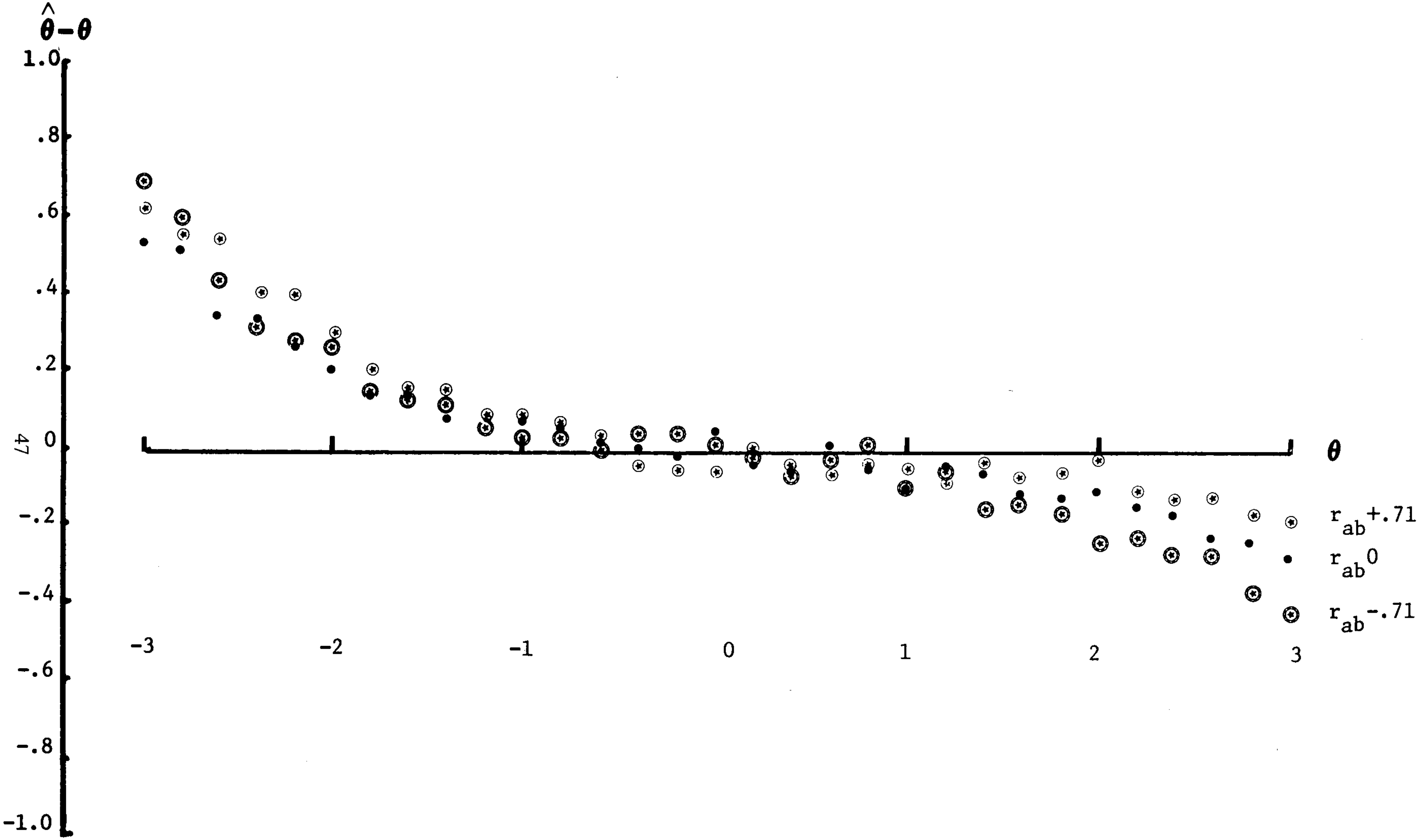


Figure 6. Mean error of estimate $\hat{\theta}-\theta$) at thirty-one ability points (θ) for the simulated Bayesian sequential test under three item pool configurations.

TABLE 5

Estimated Value of the Derivative $\frac{\partial \hat{\theta}}{\partial \theta}$, Conditional Standard Deviation $\sigma_{\hat{\theta}|\theta}$ and Value of the Information Function $I_{\hat{\theta}}(\theta)$ for Three Item Pool Configurations, at 31 Ability Levels (θ)

	Item Pool Configuration								
	r_{ab}.71			r_{ab}0			r_{ab}-.71		
θ	$\frac{\partial \hat{\theta}}{\partial \theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$	$\frac{\partial \hat{\theta}}{\partial \theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$	$\frac{\partial \hat{\theta}}{\partial \theta}$	$\sigma_{\hat{\theta}\|\theta}$	$I_{\hat{\theta}}(\theta)$
-3.0	.523	.307	2.90	.588	.336	2.58	.450	.353	1.63
-2.8	.566	.353	2.57	.629	.333	3.57	.511	.308	2.75
-2.6	.607	.328	3.42	.668	.304	4.83	.568	.279	4.14
-2.4	.645	.341	3.58	.704	.283	6.20	.621	.264	5.54
-2.2	.682	.321	4.51	.738	.294	6.31	.670	.268	6.26
-2.0	.716	.330	4.71	.770	.284	7.35	.716	.289	6.14
-1.8	.748	.324	5.33	.799	.228	12.29	.758	.289	6.87
-1.6	.778	.257	6.26	.826	.266	9.64	.796	.247	10.37
-1.4	.783	.311	6.34	.850	.265	10.29	.830	.230	13.01
-1.2	.832	.314	7.01	.872	.261	11.16	.860	.251	11.73
-1.0	.855	.278	9.46	.892	.275	10.52	.886	.235	14.21
- .8	.876	.316	7.69	.909	.278	10.70	.908	.244	13.86
- .6	.895	.283	10.00	.924	.260	12.63	.927	.244	14.44
- .4	.912	.282	10.47	.936	.288	10.57	.942	.255	14.66
- .2	.927	.308	9.06	.946	.278	11.59	.953	.284	13.96
0	.940	.305	9.50	.954	.249	14.68	.960	.257	13.96
.2	.946	.253	13.98	.959	.248	14.96	.963	.284	11.50
.4	.959	.255	14.14	.962	.281	11.72	.963	.252	14.59
.6	.965	.287	11.29	.962	.275	12.25	.958	.285	11.31
.8	.965	.269	12.86	.960	.248	15.00	.950	.276	11.85
1.0	.971	.228	18.15	.956	.250	14.62	.938	.336	7.79
1.2	.971	.228	18.13	.949	.250	14.42	.922	.294	9.84
1.4	.968	.218	19.71	.940	.272	11.94	.902	.295	9.36
1.6	.964	.246	15.35	.928	.259	12.85	.879	.301	8.52
1.8	.957	.229	17.46	.914	.292	9.81	.851	.317	7.21
2.0	.948	.263	13.00	.898	.289	9.66	.820	.296	7.67
2.2	.937	.230	16.56	.879	.260	11.43	.785	.321	5.98
2.4	.924	.210	19.35	.858	.255	11.32	.746	.294	6.44
2.6	.908	.227	16.00	.834	.270	9.55	.703	.349	4.06
2.8	.891	.258	16.69	.808	.250	10.46	.657	.332	3.91
3.0	.871	.218	16.00	.780	.279	7.82	.606	.293	4.28

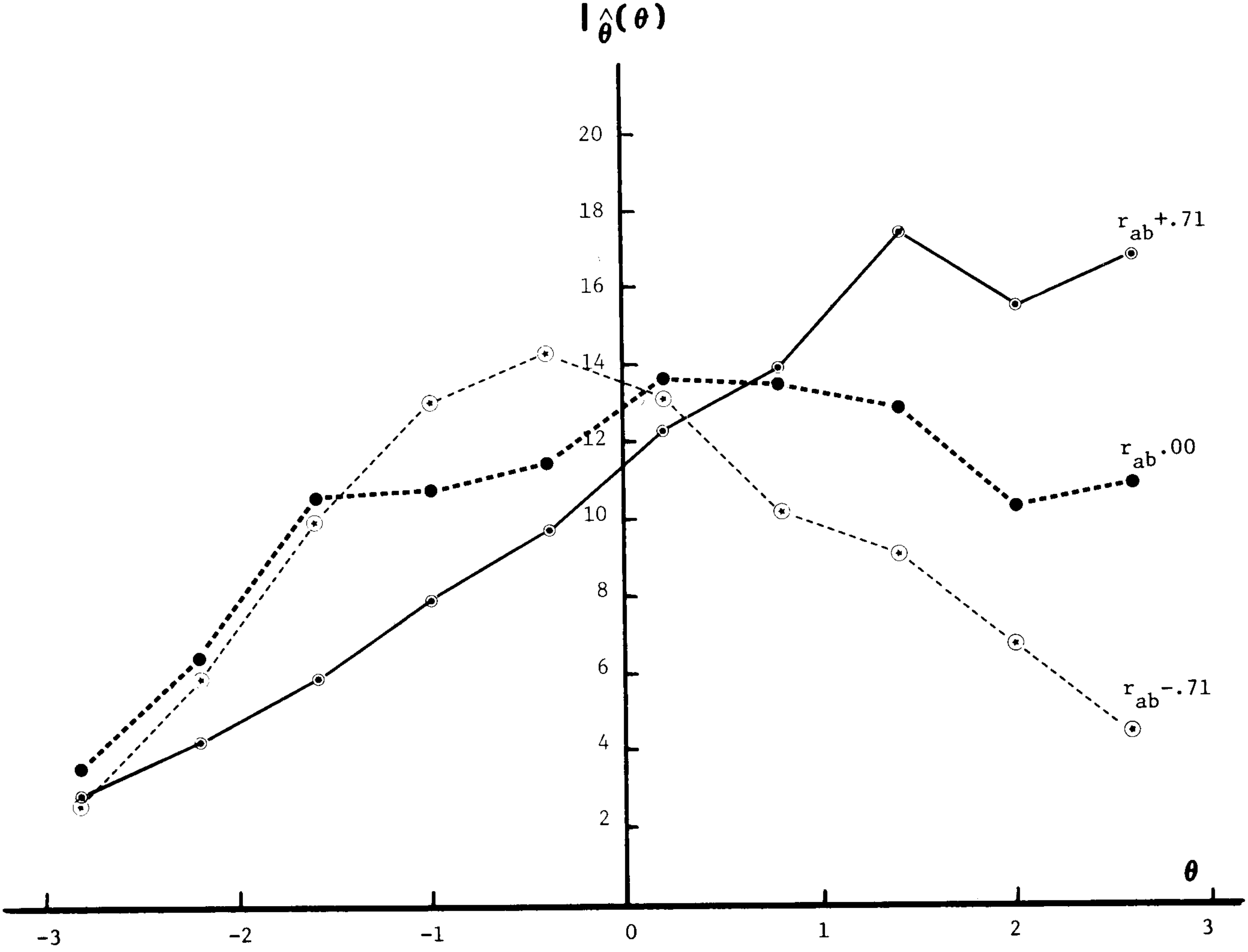


Figure 7. Smoothed curves of the information functions of the Bayesian sequential test under three different item pool difficulty-by-discrimination configurations. (see text.)

1) Under all three item pool configurations the information functions were very low in the low end of the θ distribution;

2) For $r_{ab}+.71$ the information values uniformly increased with increasing θ;

3) For $r_{ab}0$ information generally increased with θ, to about $\theta = 1.00$, then decreased somewhat;

4) For $r_{ab}-.71$ information increased sharply with θ, to about $\theta = 0$, then just as sharply decreased.

DISCUSSION

Study 1

Test length, or number of items required to satisfy the posterior variance termination criterion, was shown to vary inversely with item discriminatory power, a_g, when the latter is constant for all items in a given test. This result was expected, and corroborates the findings of Jensema (1972, 1974) who also pointed out that if constant item discriminatory powers were available it would be possible to predict the validity of the trait estimates from the number of items administered, and conversely to estimate the number of items required to achieve any given validity level.

In the no-guessing and uncorrected guessing conditions (that is, in tests which assume no guessing) the test length was constant for any fixed a_g value. This result would not be likely to occur with a finite pool of items due to the inevitability of imperfect θ-with-item-difficulty matches. That is, with a finite item pool some variance in test length would likely occur even if all items had equal discrimination parameters. The fact that there was no variance in test length (within any given discrimination level) with the perfect item pool indicates that any variance in test length in a real, constant-discrimination, no-guessing test must be due solely to inadequacies in the distribution of item difficulty parameters in the finite item pool.

These results are pertinent to the use of Rasch-model ability estimation in an adaptive testing situation. Except for the specification of the item characteristic function, the Rasch model is conceptually identical with the no-guessing model used in Study 1. Within each test, item discrimination parameters were constant (as the Rasch model assumes) and no-guessing was assumed. Thus the major difference between this portion of Study 1 and a Rasch model simulation would be in the definition of the item response model. We assumed a one-parameter normal ogive response model, whereas the Rasch model uses a one-parameter logistic one (Birnbaum, 1968, p. 402). As Birnbaum (1968, p. 399) has pointed out, the two response models are very similar. Thus, the results of Study 1 for the no-guessing condition should be generalizable to adaptive tests based on the Rasch model.

In the corrected guessing condition (Figure 3) there was some variance in test length for all a_g values (except $a_g = .50$, where no testees terminated in fewer than 100 items). For all a_g levels above .50, test length θ correlated strongly and positively with the trait estimate $\hat{\theta}$ (Table 3). The test length $-\hat{\theta}$ correlation $r_{\hat{\theta}k}$ equalled or exceeded .80 for all a_g values above .6. The correlation $r_{\theta k}$ between test length and ability θ was of similar magnitude but always smaller than $r_{\hat{\theta}k}$. It seems obvious that for the case of constant item discrimination and non-zero guessing there is a systematic relationship between ability θ or test score $\hat{\theta}$ and number of items administered. Examination of the partial correlations, however, shows that $r_{\theta k}$ vanishes when θ is statistically controlled for. For instance, for $a_g = 1.0$ we observed $r_{\theta k} = .81$, $r_{\hat{\theta}k} = .83$, $r_{\theta\hat{\theta}} = .98$. Controlling for $\hat{\theta}$ and θ, respectively, yields the following partial correlations:

$$r_{\theta k.\hat{\theta}} = -.03$$

$$r_{\hat{\theta}k.\theta} = .31$$

Analysis of the corresponding partial correlations for the other a_g levels would yield a similar result: $r_{\theta k.\hat{\theta}}$ approximately zero, but $r_{\hat{\theta}k.\theta}$ positive and moderate. This suggests that, at least for the constant item discrimination case, the tendency for $r_{\hat{\theta}k}$ to be positive is due to some characteristic of the trait estimation method using the guessing correction.

Another observation with regard to test length has a practical application. Where the posterior variance termination criterion is to be used, it is desirable that all or nearly all examinees reach criterion (e.g., $\sigma^2_m \leqslant .0625$ or some other arbitrary value) within a reasonably small number of items. Typically (e.g., Urry, Jensema), a 30-item maximum test length has been imposed in conjunction with the posterior variance criterion. If a large number of examinees reach the 30-item limit before attaining the posterior variance criterion, the latter may lose its usefulness as a predictor of test validity. The data of Table 3 (and Figure 3) indicate that even with a "perfect" item pool, the constant item discrimination parameter must equal or exceed $a_g = 1.25$ in order to insure test termination in fewer than 30 items for the majority of examinees when guessing is a factor. Although it is difficult to generalize this finding to the case of typical finite item pools, it is reasonable to expect that test termination via the posterior variance criterion $\sigma^2_m < .0625$ will seldom occur in fewer than 30 items in Bayesian sequential tests using item pools whose mean item discrimination parameter is less than 1.25.

Errors of estimate were moderately and negatively correlated with θ in all three conditions, with the strongest correlations observed in the uncorrected guessing situation. That is, with constant item discrimination and a perfect pool of item difficulties, larger errors of estimate $(\hat{\theta}-\theta)$ tended to occur as θ decreased. This tendency can be viewed as a regression effect. As is typical with linear regression estimates for all three conditions the estimates $\hat{\theta}$ tended to be closer to the mean than the actual values θ.

The correlation $r_{\hat{\theta}e}$ between trait estimates $\hat{\theta}$ and errors $(\hat{\theta}-\theta)$ was consistently of the same sign but lower magnitude than $r_{\theta e}$, with the no guessing and corrected guessing conditions.

The mean error of estimate, or bias, was virtually zero in the no guessing condition, until a_g became large (Table 1; Figure 1). For $a_g \geqslant 1.50$ there was a tendency for positive bias to occur. Similarly, mean absolute error was quite constant until $a_g=1.50$, than became larger. In the corrected guessing condition (Table 3, Figure 3) mean absolute error was fairly constant across a_g levels, but bias was positive at low a_g values, diminished virtually to zero at intermediate levels, and began to increase steadily as a_g increased above 2.0.

Study 2

Test length. The data illustrate clearly the effect of item pool configuration on the correlation of test length with θ (or $\hat{\theta}$): The correlation is strong and its sign was opposite that of the r_{ab} correlation in the simulated item pool. (For the $r_{ab}0$ configuration there was no variance in test length, due to the arbitrary 30-item limit. The preceding three studies have shown, however, that with constant a_g, test length varies directly with θ. Presumably that relationship would hold for the $r_{ab}0$ configuration if test length was free to exceed 30 items). We have already alluded to the inverse relationship between test length and the rate of reduction in the Bayes posterior variance. Thus, it should be clear that the configuration of difficulty and discrimination parameters in the item pool, which can be roughly described by the correlation of the discrimination and difficulty parameters (r_{ab}), effectively dictates the rate of posterior variance reduction at any level of the trait θ. Furthermore, if a maximum test length is arbitrarily established (such as the 30-item limit used by us, and by Urry, 1974, and Jensema, 1972) that limit, in conjunction with the item pool configuration, may dictate regions of the θ continuum in which satisfactory convergence of the trait estimates will seldom occur.

Errors of estimate. Study 1 found very high validities of the trait estimates $\hat{\theta}$, indicating that the Bayesian sequential test is capable of ordering simulated examinees from a normal population quite well with respect to the variable, θ, underlying the item responses. Study 2 was motivated by an interest in the *accuracy* of the estimates of θ, rather than the correctness of ordering, as a function of θ itself. The data showed clearly that the Bayesian estimates behaved in a manner similar to linear regression, except at the extremes of the normal distribution ($\theta \leqslant -1.5$ and $\theta \geqslant 2.0$). Typically, linear regression underestimates the criterion variable above the mean, and overestimates it for values below the mean. Such was the case for the Bayesian sequential estimates, except that the underestimates became fairly sizeable (around .20) on the average for $\theta > 2.0$, and overestimates became severe (larger than .5) in the lower levels of the trait. Furthermore, it was shown that the behavior of the trait estimates varies as a function of the item pool configuration. Thus, by controlling the item pool configuration for a live-testing item pool it should be possible to control the accuracy of the Bayesian test scores as estimators of the actual trait level of the examinees. Other alternatives may prove useful in this regard. Some of these will be discussed below.

Information. For the configuration $r_{ab}+.71$, the information of the trait estimates appears to increase linearly with θ, at least in the interval $[-3.0 \leqslant \theta \leqslant 3.0]$. This is what we might expect, since item discrimination increased with θ in this configuration. Note (Table 4) that mean test length in this configuration was 30 items for $-3 \leqslant \theta \leqslant .6$, and then decreased linearly with for $\theta < .6$, reaching a mean of 23 items at $\theta = 3.0$.

For the $r_{ab}0$ configuration the information function appeared to take the shape of an inverted (and rather asymmetric) shallow dish, with maximal information attained in the interval $[0 \leqslant \theta \leqslant 1.5]$. This should approximate, at least in its form, the information structure resulting from applying the Bayesian sequential test with a real item pool whose configuration is based on Urry's (1974) prescription. It should be apparent that some efficiency of measurement will be lost in the extremes of the θ distribution, especially in the lower extremes. Note that for these data, test length was a constant 30 items at all levels.

For the $r_{ab}-.71$ configuration the information curve does not take the shape one would assume intuitively. From knowledge of the distribution of the discrimination parameters it would seem that the curve should mirror that of the $r_{ab}+.71$ information but with maximal information at $\theta=-3.0$. Instead it rather emphatically takes the convex form. The test is maximally efficient in the interval $[-1 \leqslant \theta \leqslant 0]$, and rapidly loses efficiency elsewhere. This is a remarkably different result from what one would expect. The highest item discrimination parameters were available at the low end of the θ scale, yet information was as low there $[-2 < \theta < -1.5]$ as it was where the lowest item discrimination values occurrcd $[1.5 < \theta < 3.0]$. The low levels of information in the low θ region are due in part to the small number of items administered there. As Table 4 reveals, the posterior variance termination criterion resulted in mean test length of 14 items at $\theta=-3.0$; 17 items at $\theta=-2.0$; 22 items at $\theta=-1.0$ The information values obtained with these test lengths could be adjusted statistically to estimate the information values for constant 30 item test length. Such an adjustment would still show an efficiency loss at $\theta < -2.0$ for this item pool configuration, despite the high average item discrimination in that region. We will address this problem further in the discussion to follow.

Implications. These results were obtained by simulating a "perfect" item pool; i.e., a pool in which unlimited numbers of items of any difficulty level were available. This should result in data, which, within the limits of sampling error, approximate the best possible results obtainable using the sequential testing procedure as specified by Owen (1969), under the conditions studied.

We have found, as did Urry (1971, 1974) and Jensema (1972, 1974) before us, that the procedure has the potential to yield trait estimates having very high validities with great economy in test length, provided that highly discriminating test items, rectangularly distributed on difficulty, consitute the item pool. We have also found that there may be a tendency of the method to overestimate group mean trait level, when item discrimination parameters are very high, even when the trait estimation model exactly conforms to the item response model. When the estimation model is not congruent with the item response model (as in the uncorrected guessing condition of study 1) we have found that rather sizable bias of estimate may occur, accompanied by diminished validity.

Lord (1970, p. 152) made the point that evaluating a tailored test by means of a group statistic (such as our validity coefficient $r_{\theta\hat{\theta}}$) presumes some knowledge of the group's distribution on the trait being measured, and ignores information relevant to the accuracy of trait estimates at any one level of the trait. The validity of the Bayesian sequential test trait estimates was, as we have seen, quite high under the conditions used in our simulation studies. The accuracy of the estimates was also favorable in what corresponds to the middle ranges of a normal distribution on θ, but was found to be less favorable in the extremes, especially the lower extreme. Similarly, the information functions of the trait estimates showed that the effectiveness of measurement under the Bayesian tailoring procedure varied systematically as a function of the configuration of the item parameters constituting the item pool, but in all three configurations measurement effectiveness was very low in the low ranges of the trait.

The observed loss of accuracy and information in the extremes of the "typical" range of θ are disturbing; since the advantage of tailored testing over conventional testing is the former's supposed potential for superior measurement accuracy and effectiveness in those extremes. From our data it is apparent that with the exception of the r_{ab}+.71 configuration, the sequential test scores are behaving much like conventional test scores, at least in terms of the shapes of their information functions. And even for the r_{ab}-.71 configuration measurement effectiveness was relatively poor in the lower extremes of θ. The utility of the Bayesian adaptive testing strategy may be diminished considerably by results like those reported for Study 2, if they prove to be general.

The problems revealed in Study 2 (of bias non-linear in θ, and of convex information structures of the trait estimates) have causes which may be amenable to improvement. At the heart of the problem is the effect of guessing, which generally operates to reduce measurement efficiency at all trait levels, and especially at low trait levels. Also at the core of the problem is the Bayesian procedure itself. As we have pointed out earlier, the Bayesian trait estimates behave like regression estimates. Extreme values of θ are systematically regressed toward the initial prior estimate: the assumption of a normal prior distribution of θ ensures this tendency. Now, the more extreme θ is for any individual, the larger will be the regression effect, on the average. Recall that the item selection procedure selects an item with difficulty b_g somewhat easier than the current θ estimate. But for high θ the current estimate is almost always too low. Hence the difficulty of the selected item will almost always be too easy for extremely able examinees. Cumulated over, say 30 items, the effects of this inappropriate item selection will be several:

1) mean proportion correct will tend to increase as a function of θ, despite the explicit attempt of the tailoring procedure to make it constant at all levels of θ;

2) θ will tend to be underestimated for high θ due to the inappropriate difficulty of the test items administered;

3) information loss will occur at high θ due to the shallowing slope of the regression of $\hat{\theta}$ or θ.

For low θ the initial prior is an overestimate. Hence, the first item selected will generally be too difficult $[(b_g-\theta)>0]$, yet the examinee has a non-zero chance of answering it correctly. A correct answer, of course, will cause an increase of $\hat{\theta}$ and thus result in another inappropriate choice of item difficulty. Furthermore, as Samejima (1973) has shown, there may actually be negative information in a correct response to an item whose difficulty b_g exceeds an examinee's actual trait level θ by a fairly small increment, when guessing is a factor. We suggest that examinees in the low extremes of θ are rather consistently being administered overly difficult items $[(b_g-\theta)>0]$ with several systematic results:

1) mean proportion correct tends to decrease with θ despite the tailoring process;

2) posterior variance reduction tends to be more rapid for individuals of low trait levels, due largely to their sub-optimal proportion of correct responses, resulting in shorter mean test length;

3) the shorter the test length, the less opportunity the Bayesian estimation procedure has to converge to extreme trait level estimates;

4) non-convergence combines with negative information in some correct responses to diminish severely the effectiveness of measurement in the low regions of the trait.

Some of the conclusions just stated are speculative. Specifically, we have not looked at proportion correct as a function of θ, nor at the quantity $(b_g-\theta)$, both of which bear on the appropriateness of the tailoring process. Future simulation studies will be necessary to examine these variables.

One goal of adaptive testing should be to achieve a constant high level of measurement effectiveness at all levels of θ. This desideratum is equivalent to a high, horizontal information function. We have found that the Bayesian sequential test failed to achieve this goal despite an unrealistically favorable set of circumstances: the perfect item pool, errorfree item parameters, and a scoring model perfectly congruent with the item response model. We have attributed the shortcomings of the Bayesian trait estimates to the regression-like tendency of the sequential estimates themselves, which in turn result in inappropriate item selection for individuals whose trait levels are extremely high or low.

There are at least two methods of ameliorating this problem, both of which should, to some extent, lessen the bias of estimate at the extremes and improve the information structure of the trait estimates. The first method involves the assumption of a rectangular rather than a normal prior distribution of θ. The second method would involve replacing the present item selection procedure with a mechanical branching procedure which would be less sensitive to large errors in the current trait estimate in its choice of the next item to administer. Needless to say, both of these alternatives do considerable violence to Owen's elegant procedure.

If the practitioner is committed to the procedure as it was originally proposed, it would seem thât the best course of action would be to take great care in assembling the item pool, and to administer a constant number of items (say 30) to each examinee. If no strong commitment to Owen's procedure is involved, the practitioner may be well advised to use another adaptive strategy, such as Weiss' stradaptive test (Weiss, 1974), Lord's (1974) maximum likelihood procedure, or a similar procedure being investigated by Samejima (1975). Systematic investigation of some of these strategies, which will permit them to be compared with the Bayesian sequential test, are currently in progress.

REFERENCES

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

Jensema, C. J. An application of latent trait mental test theory to the Washington Pre-college Testing Program. Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766.

Lord, F. M. Some test theory for tailored testing. In Holtzman, W. H. (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970 (Chapter 8).

Lord, F. M. A broad-range test of verbal ability. Research Bulletin 75-5. Princeton, N. J.: Educational Testing Service, 1975.

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

McBride, J. R. & Weiss, D. J. Simulation studies of Bayesian adaptive ability testing, 1975 a. (In preparation.)

McBride, J. R. & Weiss, D. J. An empirical study of Bayesian computerized testing, 1975b. (In preparation.)

Owen, R. J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, N. J.: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, in press.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika* 1973,*38*(2), 221-233.

Samejima, F. Behavior of the maximum likelihood estimate in a simulated tailored testing situation. Paper presented at the meeting of the Psychometric Society, Iowa City. April 1975.

Urry, V. W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: Bureau of Testing, University of Washington, 1971.

Urry, V. W. Computer assisted testing: the calibration and evaluation of the verbal ability bank. Technical study 74-3. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974.

Weiss, D. J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1973.

Wood, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation. University of Chicago, 1971.

AN EMPIRICAL INVESTIGATION OF WEISS' STRADAPTIVE TESTING MODEL

BRIAN K. WATERS
U.S. Air Force Human Resources Laboratory

This study[1] investigated the validity and utility of the stratified adaptive ("stradaptive") computerized testing model proposed by Weiss and colleagues in the Psychometric Methods Program, University of Minnesota. Weiss and his associates have reported the theoretical development of the stradaptive model (Weiss, 1973; DeWitt and Weiss, 1974; McBride and Weiss, 1974) including some examples of individual results. To date, no full empirical studies of the model have been published.

The Stradaptive Testing Model

Lord's theoretical analysis of adaptive testing versus conventional testing makes one point very clear: a peaked test provides more precise measurement than an adaptive test of the same length *when the testee's ability is at the point at which the conventional test is peaked.* At some point on the ability continuum, generally beyond ± .5 standard deviations from the mean, the adaptive test requires fewer items for comparable measurement efficiency.

Lord suggests that an "ideal" testing strategy would present a sample of items to each subject comprising a peaked test with a .50 probability of a correct answer for examinees of the particular subject's true ability (P_c = .50). The catch, of course, is that the true ability of the subject is unknown; the estimation of which is, in fact, the desired outcome of the measurement procedure.

Traditionally, this problem has been circumvented by peaking the test at P_c = .50 for the hypothetical *average* ability level subject. This procedure worked well for examinees near the center of the ability continuum, but less efficiently near the extremes.

Weiss' stradaptive model extends the Binet rationale to computer-based ability measurement. A large item pool is necessary, with item parameter estimates based upon a large sample of subjects from the same population as potential examinees. Items are scaled into peaked levels (strata) according to item difficulty. A subject's initial item is based upon a previously obtained ability estimate or the subject's own estimation of his ability on the dimension being assessed.

Figure 1 depicts a nine-strata distribution of items in a hypothetical stradaptive item pool.

As in the Binet, the subject's basal and ceiling strata are defined, with testing ceasing when the ceiling stratum has been determined. A subject's score is a function of the difficulty of the items answered correctly, utilizing various scoring strategies (Weiss, 1973).

The Item Bank

Verbal analogy test items were used in this study selected from the SCAT Series II.[2] This test series provided a single-format, unidimensional test with extensively-normed item parameter estimates. The item format was easily stored in a computer item file, being short and standard for all 244 items.

Item pool data received from Educational Testing Service contained five 50-item verbal analogy tests, Forms 1A, 1B, 1C, 2A and 2B of the SCAT Series II examinations. These tests had been nationally normed on a sample 3133 twelfth grade students in October 1966. P–values and biserial correlations on 249 items were provided by ETS. These values were transformed into normal ogive item parameters.

Table 1 shows the actual distribution of items used in this experiment. The final pool included 244 items grouped into 9 strata according to normal ogive item difficulty parameters as shown in Table 1.

The nine strata in Table 1 are essentially nine peaked tests, varying in average difficulty from −2.12 to +1.91. Stratum 9, the most difficult peaked test, for example, was composed of 19 items ranging from $b_g = 1.27$ to b_g=3.68. In this study, items were randomly ordered within strata, unlike in Weiss' model, in order to permit an alternate-forms reliability coefficient to be calculated for stradaptive examinees. As is typical in educational and psychological research, the concentration of more difficult items contains the lower discrimination values. A correlation between b_g and a_g of −.31 reflects this problem.

Subject Pool. One hundred and two incoming freshmen to Florida State University were tested in late July 1974. Ninety-nine of the subjects had Florida Twelfth Grade

[1] This paper is based on the author's doctoral dissertation conducted at Florida State University under the direction of Dr. Howard W. Stoker. Requests for copies of the dissertation should be sent to the author c/o AFHRL/FT, Williams AFB, AZ 85224.

[2] Test materials from SCAT Series II Verbal Ability tests were adapted and used with the permission of Education Testing Service. The author of this paper gratefully acknowledges the help of ETS in the pursuit of this research.

(12V) Verbal Scores or 12V estimates derived from ACT or CEEB verbal scores to serve as criteria for the validity investigation of the stradaptive test scores.

Table 2 depicts linear vs stradaptive group test statistics on the 12V scores.

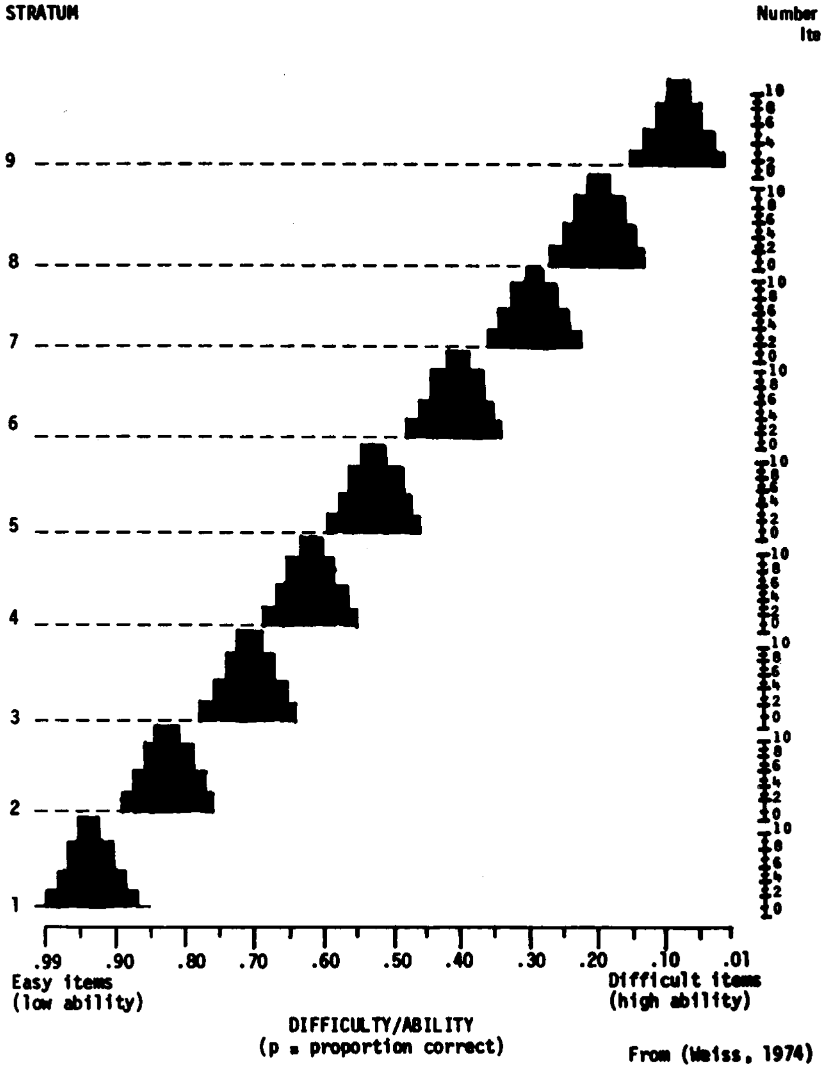
As can be seen in Table 2, the random assignment of subjects to linear or stradaptive testing groups did a good job in equating the groups on the ability continuum as presented.

Testing continued until a subject's ceiling stratum was identified. for this study, the ceiling stratum was defined as the lowest stratum in which 25% or less of the items measured by the Florida 12th Grade Verbal test.

Since SCAT-V published results had shown significantly different difficulty levels between the five forms, linear subtest scores were normalized within their separate distributions and then pooled into a linear total score distribution for comparison with stradaptive results.

CRT Testing

A computer program described by DeWitt and Weiss (1973) was adapted to fit the FSU Control Data Corporation 6500 computer.



From (Weiss, 1974)

Figure 1. Distribution of items, by difficulty level, in a Stradaptive Test

TABLE 1

Item Difficulties (b) and Discriminations (a), Based on Normal Ogive Parameter Estimates, for the Stradaptive Test Item Pool

									Stratum									
			(easy)										(difficult)					
	1		2		3		4		5		6		7		8		9	
Item Difficulties																		
High	-1.94		-1.46		- .90		- .49		- .10		.25		.67		1.34		3.68	
Low	-3.57		-1.91		-1.40		- .88		- .44		- .10		.27		.71		1.27	
Mean	-2.12		-1.68		-1.13		- .68		- .25		.04		.44		.95		1.91	
No. of Items	20		26		33		39		31		28		26		22		19	
Item Number Within Stratum	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a
1	-2.08	.48	-1.87	.50	- .90	.56	-.62	.79	- .19	.29	.20	.59	.38	.53	1.25	.46	1.76	.49
2	-1.97	.45	-1.74	.69	-1.05	.83	-.49	.69	- .33	.41	.25	.59	.31	.34	.76	.64	1.69	.44
3	-2.07	.42	-1.70	.95	-1.34	.42	-.72	.77	- .11	.50	.00	.44	.43	.53	1.19	.56	1.61	.44
4	-2.27	.64	-1.91	.52	-1.11	1.17	-.65	.73	- .17	.50	.24	.34	.63	.39	.81	.45	2.91	.49
5	-1.97	.86	-1.50	.77	-1.39	.63	-.88	.49	- .11	.53	.09	.69	.65	.66	1.13	.36	3.69	.28
6	-2.17	.36	-1.79	.59	- .92	.50	-.49	.61	- .16	.33	-.10	.61	.34	.49	.87	.34	1.57	.29
7	-2.31	.41	-1.47	.86	-1.06	.42	-.80	.50	- .16	.83	.09	.71	.30	.71	.71	.48	1.60	.33
8	-2.03	.41	-1.83	.55	-1.31	.44	-.69	.75	- .11	.52	.12	.81	.59	.53	.88	.53	1.34	.42
9	-2.13	.48	-1.68	.58	-1.22	.95	-.55	.52	- .44	.69	.11	.52	.28	.64	.88	.49	1.83	.52
10	-3.57	.30	-1.52	.93	-1.08	.52	-.80	.55	- .42	.55	.00	.39	.38	.52	.79	.61	1.27	.77
11	-2.03	.50	-1.69	.83	-1.19	.79	-.57	.33	- .21	.39	.00	.41	.29	.61	1.24	.30	2.29	.44
12	-2.63	.39	-1.69	.41	- .95	1.01	-.84	.68	- .35	.59	.21	.53	.62	.56	.71	.71	1.33	.33
13	-1.95	.25	-1.65	.71	-1.37	.53	-.86	.83	- .24	.79	.13	.68	.53	.55	.91	.26	1.91	.40
14	-1.95	.56	-1.56	.64	-1.31	.64	-.76	.59	- .10	.55	-.08	.77	.29	.37	1.06	.49	1.27	.42
15	-2.31	.63	-1.90	.69	-1.40	.75	-.54	.46	- .42	.75	.13	.44	.27	.68	1.24	.33	1.91	.27
16	-2.50	.53	-1.51	.88	- .90	.36	-.53	.73	- .41	.66	.00	.71	.56	.45	1.01	.56	2.94	.25
17	-2.03	.50	-1.88	.59	-1.04	.68	-.83	.58	- .16	.83	-.05	.56	.67	.46	.75	.79	1.94	.41
18	-2.36	.61	-1.83	.90	- .97	.81	-.51	.58	- .30	.58	.13	.44	.40	.59	1.34	.37	2.13	.27
19	-1.95	.81	-1.80	.36	-1.09	.68	-.62	.79	- .31	.34	.14	.66	.32	.53	.95	.25	1.33	.37
20	-2.03	.71	-1.55	.61	- .91	.77	-.86	.55	- .31	.45	.05	.64	.30	.73	.75	.66		
21			-1.65	.45	-1.02	.75	-.64	.68	- .18	.68	-.91*	.97	.29	.48	.79	.46		
22			-1.78	.68	-1.18	.46	-.85	.46	- .33	.64	-.06	.50	.66	.64	.94	.53		
23			-1.50	.77	-1.35	.45	-.59	.77	- .35	.69	.12	.77	.37	.66				
24			-1.46	.63	-1.17	.58	-.53	.41	- .18	.48	.06	.50	.56	.70				
25			-1.46	.49	-1.07	.27	-.65	.66	- .44	.52	.10	.55	.50	.68				
26			-1.90	.79	- .95	.66	-.75	.73	- .16	.81	.00	.45	.56	.39				
27					-1.36	.98	-.54	.88	- .23	.49	-.04	.88						
28					-1.27	.71	-.60	1.07	- .19	.44	.07	.36						
29					-1.39	.88	-.74	.61	- .37	.79								
30					- .90	.71	-.61	.64	- .14	.66								
31					-1.30	.69	-.83	.81	- .18	.48								
32					-1.38	.36	-.75	.73										
33					-1.21	.45	-.60	.59										
34							-.88	81										
35							-.77	.48										
36							-.49	.33										
37							-.65	.33										
38							-.76	.40										
39							-.73	.83										

*This item was misassigned to stratum 6 rather than 3. Fortunately, no subjects reached the item in the Stradaptive Pool.

TABLE 2

Comparison of Distributions of Linear and Stradaptive Group Florida 12th Grade Verbal Scores

GROUP	# SUBJECT	MEAN	STD DEV	STD ERR	KURTOSIS	SKEWNESS
LINEAR	46	33.26	5.30	.855	.44	.70
STRADAPTIVE	53	34.06	6.12	.841	.36	-.03

P_r (μ lin = μ str)= > .05

P_r (σ^2 lin = σ^2 str) = > .05

Testing Sequence. The subjects estimated their ability using the procedures described in DeWitt and Weiss. The first item that the stradaptive subject received was the first item in the stratum commensurate with his ability estimate. The subject was then branched to the first item in the next higher or lower stratum depending upon whether the initial response was correct or incorrect. If the subject entered a question mark (?), the next item in the same stratum was presented.

Testing continued until a subject's ceiling stratum was identified. For this study, the ceiling stratum was defined as the lowest stratum in which 25% or less of the items attempted were answered correctly, with a constraint that at least 5 items be taken in the ceiling stratum. The 25% figure reflects the probability of getting an item right by random guessing on a 4-option multiple choice test. Once a subject's ceiling stratum was defined, the program looped back to the examinee's ability estimate stratum and commenced a second stradaptive test with item selection continuing down the item matrix from where the first test ended. Since items were randomly positioned within each stratum, parallel, alternate forms were taken by all subjects who reached termination criterion on the first test.

A maximum of 120 items per subject was established, as pre-study trial testing suggested that subjects became saturated beyond this point.

Termination Rules. Weiss had two versions of his stradaptive testing computer program. Version one, which was used in this study, presented another item in the same stratum when a subject skipped an item.

The author of this study was unaware of the existence of the second branching strategy program prior to completion of data collection. However, Weiss' program procedure of ignoring skipped items in determining test termination was questioned. It appeared that valuable information was being lost when the Weiss procedure was followed.

It was reasonable to expect that a subject would omit an item *only* when he felt he had no real knowledge of the correct answer. Thus, investigation of test termination based upon omits counted as wrong answers was judged appropriate.

Weiss had set 5 items in the ceiling stratum as the minimum constraint upon termination. A secondary goal of the present study was to determine what effect the reduction of this constraint to 4 would have upon the effectiveness of the stradaptive strategy.

These two questions of the handling of omits and the variation in the constraint on the termination of testing created the following three methods for comparisons:

Termination Method 1:
Omits ignored/constraint = 5 items
Termination Method 2:
Omits = wrong/constraint = 5 items
Termination Method 3:
Omits = wrong/constraint = 4 items

Data was collected using Termination Method 1 and then rescored using Methods 2 and 3. This was possible since no indication of the termination of the first test was given to the subject and since items were randomly ordered within strata. Once test termination was reached using Termination Method 2 or 3, the next item taken by the subject in his entry point stratum acted as the start of a parallel-forms test under the termination rule used.

Of course, Method 2 required fewer items than Method 1 and Method 3 considerably fewer than Method 2. The thrust of this investigation, then, was to determine the relative efficiency of the three methods in comparison with one another and with linear testing after equalizing test length using the Spearman-Brown prophecy formula.

Stradaptive Test Output. Figure 2 provides an example of a stradaptive test report from this experiment. A "+" next to an item indicates a correct response; a "–", an incorrect response, and "?" shows that the subject omitted the item.

The examinee in Figure 2 estimated her ability as "5." Hence, her first item was the first item in the 5th stratum. She correctly answered this question but missed her second item, and after responding somewhat inconsistently for the first nine items, "settled down" with a very constant pattern for items 10 through 19 when she reached stopping rule criterion and her first test terminated.

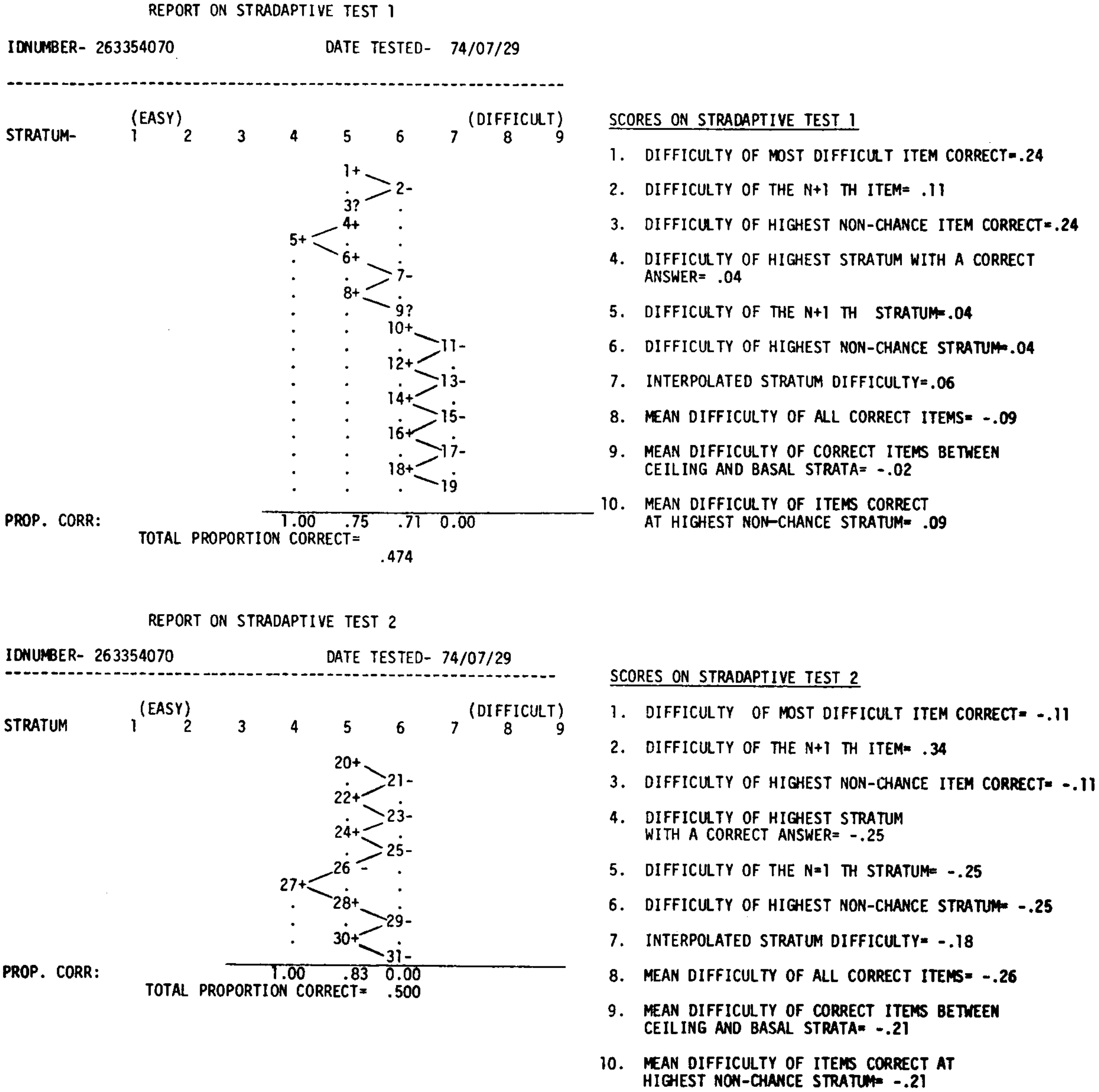
The testing algorithm then selected the 6th item in stratum 5 (her ability estimate) to commence her second test. (The subject was totally unaware of this occurrence as no noticeable time delay occurred between her 19th and 20th items).

At the conclusion of her 31st item, this subject reached termination criterion for her second test, was thanked for her help in this research project, and given her score of 15 correct answers out of 31 questions with a percentage correct of 48.4%.

The scores for this subject are shown for both tests. The interested reader may gain a more thorough understanding of the scoring methods used in this model by tracing this subject's ability estimate scores through Table 1.

RESULTS AND DISCUSSION

Test theory suggests that measurement efficiency is maximized at $P_c = .50$ for a given test group. It was hypothesized that the stradaptive test strategy would more nearly approach this standard than the conventional linear test, indicating an improved selection of items for the stradaptive subject. Table 3 shows the result of this comparison. It clearly indicates significantly different distributions of test difficulty. The stradaptive test was far more difficult than the linear test, with a smaller variance.



```
              REPORT ON STRADAPTIVE TEST 1

IDNUMBER- 263354070               DATE TESTED-  74/07/29
----------------------------------------------------------

              (EASY)                          (DIFFICULT)
STRATUM-       1    2    3    4    5    6    7    8    9
                                   1+
                                   .    2-
                                   3?   .
                                   4+   .
                              5+   .    .
                              .    6+   .
                              .    .    7-
                              .    8+   .
                              .    .    9?
                              .    .    10+
                              .    .    .    11-
                              .    .    12+  .
                              .    .    .    13-
                              .    .    14+  .
                              .    .    .    15-
                              .    .    16+  .
                              .    .    .    17-
                              .    .    18+  .
                              .    .    .    19
PROP. CORR:                   1.00 .75  .71  0.00
              TOTAL PROPORTION CORRECT=
                                        .474
```

```
SCORES ON STRADAPTIVE TEST 1

 1.  DIFFICULTY OF MOST DIFFICULT ITEM CORRECT=.24
 2.  DIFFICULTY OF THE N+1 TH ITEM= .11
 3.  DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT=.24
 4.  DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT
     ANSWER= .04
 5.  DIFFICULTY OF THE N+1 TH  STRATUM=.04
 6.  DIFFICULTY OF HIGHEST NON-CHANCE STRATUM=.04
 7.  INTERPOLATED STRATUM DIFFICULTY=.06
 8.  MEAN DIFFICULTY OF ALL CORRECT ITEMS= -.09
 9.  MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN
     CEILING AND BASAL STRATA= -.02
10.  MEAN DIFFICULTY OF ITEMS CORRECT
     AT HIGHEST NON-CHANCE STRATUM= .09
```

```
              REPORT ON STRADAPTIVE TEST 2

IDNUMBER- 263354070               DATE TESTED- 74/07/29
----------------------------------------------------------

              (EASY)                          (DIFFICULT)
STRATUM        1    2    3    4    5    6    7    8    9
                                   20+
                                   .    21-
                                   22+  .
                                   .    23-
                                   24+  .
                                   .    25-
                                   26 - .
                              27+  .    .
                              .    28+  .
                              .    .    29-
                              .    30+  .
                                        31-
PROP. CORR:                   1.00 .83  0.00
              TOTAL PROPORTION CORRECT=  .500
```

```
SCORES ON STRADAPTIVE TEST 2

 1.  DIFFICULTY  OF MOST DIFFICULT ITEM CORRECT= -.11
 2.  DIFFICULTY OF THE N+1 TH ITEM= .34
 3.  DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= -.11
 4.  DIFFICULTY OF HIGHEST STRATUM
     WITH A CORRECT ANSWER= -.25
 5.  DIFFICULTY OF THE N=1 TH STRATUM= -.25
 6.  DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= -.25
 7.  INTERPOLATED STRATUM DIFFICULTY= -.18
 8.  MEAN DIFFICULTY OF ALL CORRECT ITEMS= -.26
 9.  MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN
     CEILING AND BASAL STRATA= -.21
10.  MEAN DIFFICULTY OF ITEMS CORRECT AT
     HIGHEST NON-CHANCE STRATUM= -.21
```

Figure 2. Example of stradaptive testing report.

TABLE 3

Comparison of Difficulty Distributions (P_c)
for Linear and Stradaptive Groups

GROUP	# SUBJECTS	(P_c)	STD DEV	STD ERR	KURTOSIS	SKEWNESS
LINEAR	47	.752	.123	.018	-.87	-.39
STRADAPTIVE	55	.584	.084	.011	5.14	1.97

*P_r (μ Str = μ Lin) = < .0001

**P_r (σ^2 Str = σ^2 Lin) = < .05

Linear Test Reliability. Making the standard assumptions underlying the one factor random effects analysis of variance (ANOVA), the estimated reliability coefficient of the total scores is shown in Table 4 for the linear examinees.

The internal consistency reliability estimate for the linear test was *.776* for a test of an average of 48.4 items in length. Stepped-up to 50 items via the Spearman-Brown Prophecy formula, this estimate becomes *.782.* The reported reliability of the original SCAT-V tests was .87. Using Feldt's (1965) test, *Pr* ($p_{scat} = p_{lin}$) = < .05.

It can be assumed that the difference between these reliabilities was caused by one or more of three factors:

1. Testing mode (CRT vs paper and pencil)
2. Elimination of 6 of the 250 items from the original item pool.
3. Restriction of range in subject pool for this experiment.

The latter factor most likely caused the decrease in the reliability of the test scores. The homogeneity of the subjects would yield a relatively small amount of between-person variance, which would lower the reliability estimate. It might also be mentioned that Stanley noted that intraclass item correlation is a lower bound to the reliability of the average item.

Stradaptive Total-Test Reliability. Using Stanley's (1971) procedure, it was possible to estimate the internal-consistency reliability of the person-by-item stradaptive test matrix. Of the 244 items in the stradaptive pool, only 133 items were actually presented to the subject pool in this experiment.

Weiss' Scoring Method 8 provided the only set of stradaptive test scores wherein a person's total test score was a linear function of his item scores. Hence, this scoring method was used to estimate internal-consistency reliability. Table 5 summarizes these results.

Table 6 shows the parallel-forms and *KR*-20 reliability estimates for the three termination rules used in this study. Direct comparisons can be made between the stradaptive *KR*-20 values and the .782 linear *KR*-20 estimate. According to Feldt's (1965) approximation of the distribution of *KR*-20, all of the estimates of the stradaptive test reliability are significantly ($p = < .05$) better than the linear *KR*-20 estimate *prior* to being stepped-up by the Spearman-Brown formula *Pr* (.675 < p_{20} < .858) = .95. Thus, the 19, 26, and 31 item stradaptive tests all proved more reliable than the 48 item linear test.

A comparison of the linear internal-consistency reliability coefficients (r_{tx}) and the stradaptive parallel-forms reliability estimates (r_{xx}) in Table 6 must be considered

TABLE 4
Analysis of Variance for Linear Test Person by Item Matrix

SOURCE	df	SUM OF SQUARES	MEAN SQUARES
Persons	46	37.57	.817
Error	2229	408.55	.183
Total	2275	446.12	

$r_{tx\,(lin)}$ = 1 – .183/.817 = .776

TABLE 5

Analysis of Variance of Scoring Method 8
of Stradaptive Test Person-By-Item Matrix

TERMINATION RULE	SOURCE	df	SUM OF SQUARES	MEAN SQUARES
1	Persons	54	191.941	3.555
	Error	1675	588.253	.351
	Total	1729		(r_{20} = .901)
2	Persons	54	178.870	3.312
	Error	1401	470.442	.336
	Total	1455		(r_{20} = .899)
3	Persons	54	155.841	2.886
	Error	1001	366.447	.366
	Total	1055		(r_{20} = .873)

only tentatively since they are different kinds of estimates of the true reliability. The sampling distribution of r_{xx} is known and that of r_{tx} has been approximated by Feldt (1965). Cleary & Linn (1969) compared standard errors of both indices with generated data of known p. They found the standard error of *KR*-20 to be somewhat smaller than that of the parallel-test correlation (approximately .05 vs .04 in the range of reliabilities, number of subjects, and number of items involved in this experiment.)

Linear Test Validity. The correlation of obtained linear scores with the Florida 12th Grade Scores was .477, which was significantly lower than the published SCAT-V:SAT-v correlation of .83 ($p = < .01$). As with the linear reliability, this difference most likely resulted from subject homogeneity.

Stradaptive Test Validity. The validity coefficients of the stradaptive scoring under the three termination rules is shown in Table 7. Validity was estimated by the correlation between the test scores and 12V scores. None of the validity coefficients in Table 7 were significantly different from the linear validity coefficient of .477, although stradaptive validity coefficients were consistently higher than the linear indices.

Number of Items. Table 8 shows the difference in number of items presented for the linear and the three termination methods of the stradaptive test. The consistency in average number of items presented per subject was surprisingly constant over the two parallel tests of termination methods 1 and 3. Method 2 did show a significant ($p = < .05$) drop in the average number of items on the second test, possibly due to the 60-item limit.

Item Latency. It was hypothesized that mean item latency would be higher for stradaptive subjects since they would have to "think" about each item as it was near the limit of their ability. Table 9 reflects the results of this comparison.

The hypothesis of no differences between item latencies was rejected. For the subjects in this experiment, the average stradaptive item required approximately 11% longer than the average linear item.

Testing Costs. No full cost analysis was planned for this study. However, computer costs were available for the three-day data collection. A total of $89.00 was spent over the entire period on the CDC 6500 computer. This total included core memory (CM), central processor (CP), permanent file storage (MS), data transmittal between the

TABLE 6

Comparison of Scoring Method 8 Parallel Form Reliability with KR-20 Reliability Over Three Termination Rules Stepped Up to 50 Items

		TERMINATION RULES		
		1	2	3
		(N = 12)	(N = 28)	(N = 38)
Parallel	r_{xx}(raw)	.892	.688	.732
Forms	r_{xx}(50)	.929	.806	.903
		(N = 55)	(N = 55)	(N = 55)
KR-20	p_{20}(raw)	.901	.899	.873
	p_{20}(50)	.935	.943	.947
		$\overline{K}_1$ = 31.45	$\overline{K}_2$ = 26.47	$\overline{K}_3$ = 19.2

$\overline{K}_i$ = average number of items under termination rule 1.

TABLE 7

Comparison of Validity Coefficients of Scoring Method 8 under Three Termination Rules

Termination Rule	N	r_{cx}	r_{cx}*
1	64	.536	.585
2	80	.536	.693
3	91	.499	.626

r_{cx} = Correlation between criterion measure (12V)

r_{cx}* = r_{cx} corrected for attenuation

CRT's and the computer, line printing (LP), and punch card output for 102 subjects. Data files were punched-out as they were created to assure that data would not be lost in case of hardware malfunction.

In the present study, 6 CRT's were kept on and tied to the computer continuously for 14 hours a day for 3 days in order to be ready for subject-volunteers whenever they arrived. In any institutional implementation of computer-testing outside the experimental situation, exam time would be scheduled, thus minimizing telephone line transmittal costs.

The cost of actually testing each individual came to less than 2¢ per subject for CM, CP, MS and LP time. The vast majority of the costs cited above involve 42 hours on continual tie-in to the computer, the "unnecessary" punching out of all data, and the extensive file manipulations done by the author because direct access space became critically short during data collection. The latter factor required restorage of data files from direct to indirect file space.

This cost approximation could be compared with testing costs from the reader's experience. Without trying to define conventional testing costs per se, there is still little doubt that computer-based testing costs less than conventional testing with the paper and pencil mode for any large-scale testing program.

CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH

The results of this study favor further investigation of the stradaptive testing model. The model produced consistently higher validity coefficients than conventional testing with a significant reduction in the number of items from 48 to 31, 25 and 19 for the three stradaptive termination rules investigated in the study. The internal consistency reliability for the best stradaptive scoring methods was significantly higher than the conventional KR-20 estimate, and the stradaptive parallel-forms reliability estimates were consistently higher than conventional KR-20 estimates.

No prior research was found showing a comparison of item latency data between adaptive and conventional testing modes. Results in this study clearly indicate that subjects take significantly longer to answer items adapted to their ability level, about 11% longer in the present study. This is an important result, as it indicates that future

TABLE 8

Comparison of Average Number of Items for Linear Test and Three Termination Methods of Alternate-form Stradaptive Tests

	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS
LINEAR	47	48.43	.99			
		TEST 1			TEST 2	
	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS	# SUBJECTS	AVG # ITEMS	STD DEV # ITEMS
STRADAPTIVE						
Method 1	55	31.46	18.03	38	30.92	12.54
Method 2	55	26.94	16.76	41	21.98	13.10
Method 3	55	19.20	14.06	47	18.19	11.34

TABLE 9

Comparison of Distributions of Item Latency Between Linear and Stradaptive Groups

GROUP	# ITEMS	MEAN # SEC/ITEM	STD DEV
LINEAR	2276	35.999	12.062
STRADAPTIVE	1730	40.047	13.219

$Pr\ (\mu \text{ str} = \mu \text{ lin}) = < .001$

$Pr\ (\sigma^2 \text{ str} = \sigma^2 \text{ lin}) = < .001$

research into adaptive testing of any kind should take this variable into consideration when evaluating an adaptive test strategy. The net gain of the adaptive model is a function of the testing time needed to adequately measure a subject's ability, not the number of items presented to the subject. All prior research reviewed tacitly assumed that item latency was consistent across testing strategies. This study indicated this assumption to be false.

It is recommended that future stradaptive experimental studies should consider both stradaptive branching models with a comparison of results from variation in the minimum number of items in the ceiling stratum. A comparison between variable number of stage strategies and fixed number of stage strategies is desirable.

As suggested in previous research, adaptive testing may reach "peak" efficiency at between 15 and 20 items. A comparison of stradaptive test statistics for example with $k = 10$, 15, 20 and 25 items with linear testing should investigate this hypothesis. Once the stradaptive data is collected under the variable strategy, the fixed item statistics can be determined by grading the stradaptive test after "K" items and then "starting" the subject's second test at the first item of the entry point level.

Following the same logic which led to termination of a subject's testing when five items in a row in the highest stratum had been correctly answered, the missing of five items in a row of *any* stratum should provide immediate ceiling stratum definition. The probability of this occurrence would be less than .05 for a properly normed item pool. In the case of the present study, 13 of the 55 stradaptive subjects would have terminated a stradaptive test an average of 12.1 times earlier than termination method 1, with no effect upon the other 42 subjects. The resulting stradaptive test statistics obtained from the implementation of this suggestion have not been calculated, except that the change would have reduced the average number of items presented under termination method 1 to 28.4 from 31.45 (9.7%).

Further research is recommended into adaptive testing in which both the number of stages and step-size are variable. The Bayesian strategies and Urry's model (1970) are examples of this category of adaptive measurement and further model development seems appropriate.

Research is indicated with comparisons between adaptive models as well as the traditional design of comparing adaptive methods with conventional methods. Weiss' ongoing research is beginning this work, but more is needed. The traditional comparison assumes that conventional test statistics are the criterion that an adaptive testing procedure should try to duplicate. Lord, Green, Weiss and others have argued that improved measurement of the individual at all ability levels may be hidden by the use of classical test statistics such as validity and even reliability.

One objective of this study was the attempt to estimate the degree to which the violation of the assumptions of the one-factor ANOVA model affected KR-20 reliability estimates. The assumption that items are independent of one another is clearly violated in any adaptive testing procedure. The extent of the effect this violation causes is unknown, yet most previous research in adaptive testing has only considered ANOVA KR-20 estimates.

The results from this study do not permit definitive statements on this question. Nevertheless, the three KR-20 estimates were consistently higher than the 3 parallel-forms reliabilities. Cleary & Linn's (1969) Monte Carlo study indicated that r_{20} provided better parameter estimation than parallel-forms reliability estimates, so one must question whether the higher p estimates are not the result of the dependency between items. Perhaps the only way this question can be validly investigated is through a Monte Carlo study of adaptive testing with p known and the two methods compared, for estimating $_p$.

Green (1970) stated that the computer has only begun to enter the testing business, and that as experience with computer-controlled testing grows, important changes in the technology of testing will occur. He predicted that "most of the changes lie in the future . . . in the inevitable computer conquest of testing."[3]

The stradaptive testing model appears to be one such important change.

[3] Green, B.F., Jr., In Holtzman (Ed.), p. 194.

REFERENCES

Cleary, R.A., & Linn, R.L. A note on the relative sizes of the standard errors of two reliability estimates. *Journal of Educational Measurement,* 1969, *6,* L 1, 25-27.

DeWitt, L.J., & Weiss, D.J. A computer software system for adaptive ability measurement. *Research Report, 74-1* Psychometric Methods Program, University of Minnesota, 1974.

Feldt, L.S. The approximate sampling distribution of Kuder-Richardson reliability coefficient testing. *Psychometrika,* 1965, *30,* #3, 357-370.

Green, B.F., Jr. Comments on tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance.* New York: Harper & Row, 1970.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika,* 1941, *6.* #3, 153-160.

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

McBride, J.R., & Weiss, D.J. A word knowledge item pool for adaptive ability measurement. *Research Report 74-2,* Psychometric Methods Program, University of Minnesota, 1974.

SCAT Series II, *Cooperative school and college ability tests,* Princeton: Educational Testing Service, 1967.

Stanley, J.C. Reliability. In R.I. Thorndike (Ed.) *Educational Measurement.* Washington D.C.: American Council on Education, 1971.

Urry, V.W. A Monte Carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Weiss, D.J. The stratified adaptive computerized ability test. *Research Report 73-3.* Psychometric Methods Program, Department of Psychology, University of Minnesota, September, 1973.

Weiss, D.J. Strategies of adaptive ability measurement. *Research Report 74-5.* Psychometric Methods Program, Department of Psychology, University of Minnesota, December, 1974.

USING COMPUTERIZED TESTS TO MEASURE NEW DIMENSIONS OF ABILITIES: AN EXPLORATORY STUDY

CHARLES H. CORY
Navy Personnel Research and Development Center

Because most of the research with computer-assisted test administration has been concerned with tailoring item difficulties to test takers, what appear to be important characteristics of computerized equipment for expanding dimensionality of measurement appear to have been largely ignored. Since paper-and-pencil tests are limited in terms of stimulus control and response mode, the near exclusive reliance on them for personnel selection has imposed restrictions on the types of abilities which can be measured. For example, using conventional paper-and-pencil tests, it is difficult if not impossible to present a moving stimulus, obtain measures of tracking performance, control item exposure time, record response latencies, or sequence items as a function of prior responses. Computer terminals of the type ordinarily used for programmed instruction do have these capacities.

The battery of tests developed for the present research has been especially designed to exploit the special capabilities of computer terminals for pictorial display and movement and has thus been designated the Graphic Information Processing (GRIP) series. A major interest of the research was in finding abilities which are important for on-job performance which computerized tests could measure accurately but paper-and-pencil tests could not.

As a starting point for the investigation, five traits of "real world" significance as defined by Mecham and McCormick (1969) were selected. They were Short Term Memory, Perceptual Speed, Perceptual Closure, Movement Detection, and Dealing with Concepts/Information. Empirical data on the relative importance of these attributes for work performance is available from Mecham and McCormick (1969). The study was designed to provide comparisons of computerized and paper-and-pencil tests designed to measure these attributes and to compare the computerized measures and the operational variables in terms of dimensionality and validity for job performance criteria.

The equipment used for the research consisted of the IBM 1500 system plus a cathode ray tube (CRT) display unit and a screen for film presentation linked on-line to an IBM 1130 computer. Subjects responded to visual stimuli presented on the CRT by touching a target with a light pen, or by entering a response into the typewriter keyboard. Programming was carried out in Coursewriter.

The GRIP Tests

The GRIP battery consisted of eight computer-administered tests, each designed to measure a major aspect of one or more of the five job elements.

Illustrative items from each of the GRIP tests are shown in the Appendix.

1. *Memory for Objects.* Frames showing line drawings of common objects with simple one word names were flashed on the screen at an average exposure time of about one-half second per object per frame. Number of objects per frame ranged from three to nine. After the exposure period, subjects typed in the names of all of the objects remembered.

2. *Memory for Words.* The test was identical in intention and arrangement to the Memory for Objects, but with words substituted for the pictures. Of course the object of this test was to compare the recall of words given with the recall of words generated by the candidates' recognition and labeling processes. Words were of two lengths: 3-letters and 5-letters.

3. *Visual Memory for Numbers Test.* This is a digit-span test using the same type of methodology as was used for the two preceding tests but having digits as stimuli. About 50 percent of the digits were presented sequentially and the other 50 percent were presented all at once, as a single stimulus.

4. *Comparing Figures.* The frames of this computerized measure of perceptual speed contain sets of squares or circles presented as rows, vertical columns, and right and left slant columns. Three to six stimulus pairs are shown on the screen at a time. Each stimulus has a crossbar, oriented either vertically or horizontally. Subjects are asked to record as true-false answers whether or not all crossbars of corresponding pairs in a set have the same orientations.

5. *Recognizing Objects.* For this computerized closure test partially blotted-out pictures of common objects are presented. The first presentation shows 10 percent of the area and more area is added in random increments of 10 per unit until 90 percent of the picture is exposed. Subjects enter the names of the stimuli on the keyboard.

6. *Memory for Patterns.* A test designed to measure movement detection abilities, in which patterns are formed by sequentially blinking dots. Subjects are asked to report whether or not two consecutive patterns are identical and for other items they are asked to reproduce given patterns on the CRT with a light pen.

7. *Twelve Questions.* A test which resembles the Twenty Questions game in that subjects are asked to guess the name of an object based on yes-no answers supplied by the computer to questions. It differs from Twenty Questions in that the questions are supplied in the test rather than being posed by the subject. The subject's objectives are to select those questions which provide the quickest identification of the object and to avoid questions which are redundant or useless. Scores are sums of correct responses weighted by number and characteristics of the clues received.

8. *Password.* A test which resembles the regular "Password" game in that sets of words are shown on the CRT which suggest a target word. Five separate words are shown as clues. After the first two clues and each succeeding one, the name of the object may be typed on the keyboard. Scores are sums of correct responses weighted by number of clues received.

9. *Latency and Accuracy Variables.* In addition to direct measures of the personal attributes, latency measures were computed for speed of response for the Memory for Words and the Comparing Figures tests and latency of Recognizing Objects responses (speed of closure). In addition a measure of the total extent to which the response patterns failed to duplicate the stimuli in Memory for Patterns, free response was created (PAT-ERR).

Paper-and-Pencil Experimental Tests, Biographical Variables, and Operational Tests

Together with the GRIP battery, eight paper-and-pencil tests largely drawn from the ETS Kit of Reference Tests of Cognitive Factors (French et al., 1963), and a motion picture test (Drift Direction by Gibson, 1947) composed the set of experimental tests. In addition, data for each man were obtained for two biographical variables and for the nine tests which are routinely administered and used for Navy personnel decisions.

Samples

The experimental battery was administered to students at the Navy Training Center, San Diego, during May and June of 1972. Subjects were chosen from personnel in the first two weeks of technical training for three ratings having widely varied duties. Also tested in order to increase the sample size were recruits in their final week of training who were school eligible but had not yet received post-recruit assignments.

Ten to eleven months subsequent to the testing, after the subjects had served on jobs in the Fleet for several months, supervisory ratings covering both global and job element aspects of on-job performance were collected by mailout questionnaire.

The questionnaire used was an adaptation of the Position Analysis Questionnaire, a broad-based empirically-derived instrument developed by E. J. McCormick and his associates which has been extensively used for job classification research (McCormick, Jeanneret, and Mecham, 1972). The adapted questionnaire was used to collect ratings on global performance qs well as perfomance on all of the 42 job elements which were judged by a panel of Chief Petty Officers to be relevant to the positions.

After a preliminary review of the questionnaire returns, the 22 job elements having the largest representation in the sample were selected for analysis. These 22 job elements together with the sample size for each rating for each job element are shown in Table 1. For instance, the first rating, Electrician's Mate, involved Manual Control-Non-precision Tools, Assembling-Disassembling, Hand-Arm Manipulation/Coordination, etc. In contrast the Personnelman rating required Using Written Materials, Compiling Data, Operating Keyboard Devices, Persuading/Influencing Others, etc.; and the Sonar Technician rating required Using Pictorial Materials, Using Visual Displays, Adjusting Machines/Equipment, etc. The last group consisted of personnel in undifferentiated ratings, largely apprenticeship ratings. Major aspects of the assignments of this group involved Using Spoken Verbal Communication, Manual Control Non-precision Tools, Attention to Details, Completing Work, Working with Distractions, etc.

For each rating separately, zero-order validities of the tests for supervisors' marks of the job elements were computed and comparisons were made to identify the predictability patterns of attributes for job elements and to compare the operational, experimental paper-and-pencil, and experimental computerized tests as measures of these job elements. Similar types of statistics were computed and comparisons carried out for the ratings of global job performance.

RESULTS

Most of the statistically significant zero-order validities of the operational variables were found for the 12 job elements which are shown in Table 2. The predictor variables on the left are the Armed Forces Qualification Test, GCT a test of vocabulary and verbal reasoning, ARI, a test of arithmetic reasoning, MECH, a test of basic mechanical knowledge and principles, CLER, perceptual speed, SONR and RADIO, memory for pitches and sound patterns, ETST, electrical knowledge and mathematics, SHOP, Tool Knowledge, and lastly years of education.

TABLE 1

Sample Sizes for the Twenty-Two Most Common Job Elements

Job Element	EM	PN	ST	UA
Using Written Materials		48	30	71
Using Pictorial Materials	20		32	
Using Visual Displays			35	66
Using Spoken Verbal Communication	20	52	36	92
Using Non-verbal Sounds			31	
Analyzing Information	20			
Compiling Data		49		
Manual Control-Non-precision Tools	27			80
Manual Control-Precision Tools	23			
Operating Keyboard Devices		53		
Adjusting Machines/Equipment	23		29	
Assembling-Disassembling	27			
Hand-Arm Manipulation/Coordination	22			
Hand-Ear Coordination			31	
Persuading Influencing Others		40		
Exchanging Routine Information		51		69
Unusually Good Precision			29	69
Attention to Details, Completing Work	25	51	36	102
Vigilance-Continually Changing Details	20			
Coping with Time Pressure	22	49		78
Working with Distractions		48		84
Keeping up to Date		52	30	86

TABLE 2

Significant Zero-Order Validities of the Operational Variables
for Twelve Common Job Elements

Predictor Variable	Rating	Job Element											
		Written Materials	Pictorial Materials	Visual Display	Verbal Communication	Non-Precision Tools	Adjusting Equipment	Influencing Others	Routine Information	Good Precision	Attention to Details	Working with Distractions	Up-to-Date
AFQT	ST			-33*		--		--	--			--	
GCT	PN		--	--	50*	--	--			--			36*
	UA		--		22*		--	--	27*				
ARI	ST	49**				--		--	--			--	
	UA		--		24*		--	--	41*	27*	23*		
MECH	PN		--	--		--	--	55**		--	38*		
CLER	UA	25*	--		20*		--	--		29*	30**		26*
SONR	PN		--	--		--	--			--			37*
	UA	-26*	--				--	--					-26*
RADO	EM	--		--				--	--	--	-44*	--	--
	ST	36*	41*		33*	--		--	--		37*	--	39*
	UA		--		22*		--	--					
ETST	UA		--				--	--	28*				24*
SHOP	PN		--	--	45*	--	--	42*		--			
YRED	UA		--		21*	22*	--	--	34**	31**	26**		
		Cell Ns											
	EM		15		16	21	18				19		
	PN	26		29	27			20	28		29	27	31
	ST	29	30	33	34		27			27	34		28
	UA	69		66	90	79			67	67	100		84

Note. Decimal points were omitted from validity coefficients.

Coefficients significant at $p < .05$ and $p < .01$ have been identified by single and double asterisks, respectively.

A blank cell indicates nonsignificant validity.

A double hyphen (--) indicates missing data.

Only the statistically significant coefficients are shown. The level of significance is indicated by a single underline for the five percent level and double underlines for the one percent level. Blank cells indicate non-significant validities and double dashes indicated that the *N*s were too small for validity coefficients to be computed. Rows for individual ratings which did not have any statistically significant validities have been omitted.

Operational variables were generally not effective for predicting performance on job elements in the technical ratings, and where effective did not seem to be associated with underlying relationships or constructs. For instance, the writing abilities of ST's do not appear to be logically related to scores on ARI and RADIO, but they were significantly correlated with them. Similarly, the reasons for the significant relationships between RADIO and Pictorial Materials, SHOP and Verbal Communication abilities, ARI and Communicating Routine Information, MECH and Influencing Others, and CLER with writing and verbal communication skills were not clear. Yet all of these relationships were found.

On the other hand interpretation of the significant predictor-job element validities is much more logical and consistent for the experimental tests (Table 3).

TABLE 3

Significant Zero-Order Validities of the Experimental Variables for Twelve Common Job Elements

Predictor Variable	Rating	Job Element Written Materials	Pictorial Materials	Visual Display	Verbal Communication	Non-Precision Tools	Adjusting Equipment	Influencing Others	Routine Information	Good Precision	Attention to Details	Working with Distractions	Up-to-Date
Obj. No.	EM	--		--				--	--	--	-38*	--	--
Mem. Obj.	EM	--		--			-53**	--	--	--		--	--
	UA		--			-24*	--	--				-24*	
Mem. Words	ST	42*				--		--	--			--	
	PN	29*	--	--		--	--			--			
Mem. for Nos. (V)	PN	33*	--	--	48**	--	--			--			33*
	ST	40*				--		--	--			--	
Mem. for Nos. (A)	PN	29*	--	--	42**	--	--			--			
	EM	--		--				--	--	--	-42*	--	--
Counting Numbers	UA		--				--	--			26**		
	ST					--		--	--			--	39*
Comp. Figs., Machine-paced	UA	25*	--		20*		--	--		29*	30**		26*
	PN		--	--		--	--			--			44**
Gest. Comp.	EM	--		--		-40*		--	--	--	-42*	--	--
Hidden Patterns	EM	--		--		-36*		--	--			--	--
	PN		--	--	30**	--	--	32**		--			34*
	ST		39*			--		--	--			--	35*
Rec. Objs.	ST	45*				--		--	--			--	
Mem. for Pats., Free Response	ST	40*				--		--	--			--	38*
	UA		--				--	--			24*		28**
	PN		--	--		--	--			--		36*	25*
Nonsense Syls.	EM	--		--				--	--	--	-38*	--	
Inference	PN	37*	--	--	42**	--	--	49**	37*	--			29*
Twelve Questions	PN	42**	--	--	44**	--	--	36*	55**	--		42**	31*
	ST					--		--	--			--	37*
Password	PN	33*	--	--	43*	--	--			--	30*		
	ST					--	46*	--	--			--	
WORD-LAT	EM	--		--			-41*	--	--	--		--	--
CLO-LAT	ST		-47**	-44**		--		--	--			--	-37*
PAT-ERR	UA		--				--	--			-21*		-23*
	PN		--	--		--	--			--		-37*	
FIG-LAT	PN		--	--		--	--	-36*		--			
	ST			34*		--		--	--			--	
	UA		--				--	--		31**			
Cell Ns													
	EM		20		20	27	23				25		
	PN	45			48			37	47		47	44	48
	ST	29	31	34	36		29			29	36		30
	UA	71		66	92	80			69	69	102	84	86

Note. Decimal points were omitted from the validity coefficients.

Coefficients significant at p < .05 and p < .01 have been identified by single and double asterisks, respectively.

A blank indicates nonsignificant validity.

A double hyphen (--) indicates missing data.

The first five tests are short term memory tests with the first test being the ETS Kit test of Associative Memory, the next three being computerized memory tests and the last an auditorily administered measure of digit span. Interestingly the memory tests show consistent negative correlations with job elements for Electrician's Mate and the Apprenticeship group and positive correlations for Sonar Technician and Personnelman. The correlations for PNs are for Writing and Verbal Communication Skills, two job elements for which it would be logical to expect positive correlations.

The next two tests, Counting Numbers and Comparing Figures, are respectively paper-and-pencil and computerized tests of perceptual speed. Both tests discriminate primarily for Personnelmen and the Apprenticeships ratings and the patterns of validities of the two tests were very similar.

The next three tests, together with CLO-LAT, measure perceptual closure. Gestalt Completion and Hidden Patterns were from the ETS battery, and Recognizing Objects and CLO-LAT were computerized measures. The tests have negative validities for Electrician's Mate and positive validities for Sonar Technician, with primarily visual types of elements being predicted for the latter rating.

The next test was separate parts of the computerized test designed to measure movement detection. It had significant validities for Sonar Technician and also had significant validities for Personnelmen and the Apprenticeship rating group.

Nonsense Syllogisms and Inference, measures of syllogistic reasoning from the ETS battery, and the next two tests, 12 Questions and Password, are computerized variables hypothesized to measure the same type of ability. For Personnelmen both Inference and 12 Questions were significantly related to job performance and the patterns of significant validities were very similar.

The four special variables at the bottom of Table 3 correlated with visual skills and with job elements involving accuracy and precision.

These relationships are summarized in Table 4 which shows the number of significant validities of the operational, experimental paper-and-pencil, and experimental computerized variables for the job elements in each rating in which they were present.

Major areas in which the computerized measures were useful predictors were Adjusting Equipment for Electrician's Mates, Writing and Working with Distractions for Personnelmen, and Visual Displays for Sonar Technicians. In addition computerized measures were useful supplemental predictors of communication and interpersonal relationships skills for Personnelmen. Thus, the computerized tests predicted job elements which would be expected to be central to global performance for the Personnelman and Sonar Technician ratings.

TABLE 4

Significant Zero-Order Validities of Operational and Experimental Variables for Twelve Common Job Elements

	EM			PN			ST			UA		
JOB ELEMENTS	Operating Variable	Experimental Paper-and-Pencil	Experimental Computerized	Operating Variable	Experimental Paper-and-Pencil	Experimental Computerized	Operating Variable	Experimental Paper-and-Pencil	Experimental Computerized	Operating Variable	Experimental Paper-and-Pencil	Experimental Computerized
Skill Writing				–	2	4	2	–	4	2	–	1
Pictorial Materials	–	–	–				1	1	1			
Visual Displays							1	–	2	–	–	–
Verbal Communication	–	–	–	2	3	3	1	–	–	5	–	1
Non-precision Tools	–	2	–							1	–	1
Adjusting Equipment	–	–	2				–	–	1			
Influencing Others				2	2	2						
Routine Information				–	1	1				4	–	–
Good Precision							–	–	–	3	–	2
Attention - Details	1	4	1	1	–	1	1	–	–	3	1	3
Work Distractions				–	–	3				–	–	1
Keep Up to Date				2	2	3	1	2	3	3	–	3

TABLE 5

Zero-Order Validities of Experimental Variables for Global Performance

Predictor	Validity EM (*N*=27)	 PN (*N*=54)	 ST (*N*=37)	 UA (*N*=111)
		Short Term Memory		
Object Number	–.26	.13	–.03	–.01
Memory for Objects	–.16	–.03	–.05	–.07
Memory for Words	–.33	.20	.13	.01
Memory for Numbers(V)	–.15	.20	.38*	–.01
Memory for Numbers(A)	–.15	.17	.22	.08
		Perceptual Speed		
Counting Numbers	.03	.04	$\underline{.42}$	.06
Comparing Figures, Machine-paced	.02	–.10	.07	–.06
Comparing Figures, Self-paced	.06	.07	.21	.08
		Closure		
Gestalt Completion	–.28	–.26*	.28	.06
Concealed Words	–.37*	–.14	.13	–.10
Hidden Patterns	–.04	.23	.33*	.11
Recognizing Objects	–.11	–.06	.25	–.05
		Movement Detection		
Drift Direction	–.29	.07	.02	.06
Memory for Patterns, True-false	.15	–.07	$\underline{.42}$	.07
Memory for Patterns, Free Response	.19	.21	.23	$\underline{.19}$
		Dealing with Concepts/Information		
Nonsense Syllogisms	–.30	.01	.30	–.06
Inference	.18	.19	.00	.13
Twelve Questions	–.20	$\underline{.28}$	.21	.11
Password	.08	.13	$\underline{.33}$	.04
		Special Variables		
WORD-LAT	–.24	–.06	–.05	–.11
CLO-LAT	.05	.02	–.24	–.11
FIG-LAT	–.04	.00	.02	.04
PAT-ERR	–.24	–.17	–.26	–.13

*Significant at p <.05.

Zero-order validities of the experimental variables for the global rating of job performance are shown in Table 5. Nine of the 92 validity coefficients (10 percent) were statistically significant. Of the nine, five were for computerized tests. Most of the significant validities were for Sonar Technicians. In comparison, five of 35 validities of the operational tests were statistically significant (Table 6), of which three were for the UA group.

Thus, variables in the operational battery were best for predicting global performance in apprenticeship ratings whereas those in the experimental battery were more useful for predicting performance in technical ratings, and were particularly good for predicting the performance of Sonar Technicians. Personal attributes having the highest numbers of significant validities were Movement Detection and Dealing with Concepts/Information.

TABLE 6

Zero-Order Validities of Operational Variables for Global Performance

Predictor	Rating Group			
	EM	PN	ST	UA
	(*N*=21)	(*N*=31)[a]	(*N*=35)	(*N*=109)[a]
AFQT	–.09	.15	–.12	.13
GCT	.01	.24	.11	.07
ARI	–.20	.10	.38*	.25**
MECH	.04	.23	–.04	.12
CLER	.21	–.15	.11	.19*
SONR	–.08	.15	–.08	–.03
RADO	–.06	.11	.15	.15
ETST	.16	.31	–.09	.33**
SHOP	.20	.38*	–.21	.17
YRBI	–.12	.06	.01	–.11
YRED	.11	.05	–.02	.22*

[a] Complete data were not available for some of the tests.
*Significant at $p < .05$.
**Significant at $p < .01$.

TABLE 7

Optimal Predictive Composites for Global Performance of Electrician's Mates

	R				
Predictor Set	Weight Determination	Expected Cross Validation	Predictor	Beta Weight in Final Composite	*N*
Operational Classification Test Scores	.21	.00	CLER		27
Complete Set of Experimental and Operational Variables	.37	.00	Concealed Word	–.40	27
	.49	.20	CLER	.39	
	.58	.28	Drift Direction	–.28	
	.65	.34	PAT-ERR	–.50	
	.71	.40	Memory for Words	–.40	
	.78	.53	YrBi	–.36	

Multiple regression statistics for optimal sets of the operational and experimental variables for Electrician's Mate are shown in Table 7.

The first super row shows statistics for the optimal predictive composite for the eleven operational scores and the same type of statistics for the complete battery of operational and experimental variables are shown in the second super row. The second column contains the shrunken validity coefficient for each predictor selection step. Addition of the experimental tests to the battery increased the expected cross validity substantially although the sample size is so small that these figures should be interpreted with caution. The negative beta weights for PAT-ERR and YrBi are artifacts of the direction of scaling for those variables.

The same type of finding was characteristic of the predictive composite for Personnelman (Table 8). Again the negative validity of WORD-LAT was an artifact of direction of scaling.

For Sonar Technicians (Table 9) inclusion of the experimental tests in the battery added 38 points to the shrunken multiple correlation. All of the variables selected for the complete set were measures of perceptual types of abilities.

On the other hand, the experimental variables added almost no increment to the expected cross validation for the Apprenticeship group (Table 10).

The usefulness of this type of expansion of coverage of the battery may be illustrated by reference to the abilities

TABLE 8

Optimal Predictive Composites for Global Performance of Personnelmen

Predictor Set	R		Predictor	Beta Weight in Final Composite	*N*
	Weight Determination	Expected Cross Validation			
Operational Classification Test Scores	.38	.12	SHOP	.38	30
Complete Set of Experimental and Operational Variables	.38	.12	SHOP	.22	30
	.47	.20	Gestalt Completion	−1.19	
	.64	.46	GCT	1.40	
	.71	.52	FIG-LAT	.69	
	.80	.65	WORD-LAT	−.40	
	.86	.74	Mem. for Patterns, t.f.	.37	

TABLE 9

Optimal Predictive Composites for Global Performance of Sonar Technicians

Predictor Set	R		Predictor	Beta Weight in Final Composite	*N*
	Weight Determination	Expected Cross Validation			
Operational Classification Test Scores	.38	.22	ARI	.38	37
Complete Set of Experimental and Operational Variables	.42	.28	Counting Nos.	.33	37
	.54	.40	Mem. for Patterns, t.f.	.32	
	.61	.46	Nonsense Syls.	.29	
	.66	.50	Recog. Objs.	.33	
	.73	.58	Gestalt Completion	.32	

TABLE 10

Optimal Predictive Composites for Global Performance of the Apprenticeship Group

Predictor Set	R		Predictor	Beta Weight in Final Composite	*N*
	Weight Determination	Expected Cross Validation			
Operational Classification Test Scores	.33	.28	ETST	.33	111
Complete Set of Experimental and Operational Variables	.33	.28	ETST	.33	111
	.37	.29	CLER	.21	
	.41	.32	Concealed Word	−.19	

which are being measured by the elements in each of the four predictor composites selected. Thus, for EM to the Perceptual Speed measure in the operational battery were added Closure, Movement Detection, Memory, and Accuracy of Spatial Perception from the experimental battery. For Personnelman, to the Technical Knowledge component, which provided the primary predictiveness in the operational battery, were added measures of Closure, Speed of Response and Memory from the experimental battery. For Sonar Technician, to the general mental ability component in the operational battery were added measures for the Movement Detection and Closure components from the experimental battery. And for the UA group to the measures of Technical Knowledge and Perceptual Speed from the operational battery was added a measure of Closure from the experimental battery. With the exception of the Closure measures, some of which were paper-and-pencil, most distinctive predictive validities from the experimental battery were supplied by computer-administered tests.

DISCUSSION AND CONCLUSIONS

It is clear that the experimental battery represents an increase in the breadth of abilities covered beyond those in the operational Navy battery, a considerable amount of which is attributable to the GRIP tests. Computer tests apparently provided measures of several attributes which were different from those measured by paper-and-pencil tests. Furthermore, the measurement expansions of the experimental battery served to supplement the measures of the operational battery to produce substantial increases in global validities.

The unique measurement characteristics of the GRIP tests appear to be as follows:

1. Computer administration of tests of short term recall using a variety of stimuli is feasible, and appears to offer advantages in ease of data collection and processing over paper-and-pencil tests measuring the same attributes. Furthermore, use of computerized tests to eliminate the expensive and time consuming hand scoring required by paper-and-pencil tests of short term memory would make it feasible to routinely measure these skills during personnel classification testing. Computerized measures of this attribute were found to have significant positive validities for several job elements, particularly for those dealing with communication. It is probable that use of the tests for other occupations would identify additional relationships which are useful for personnel classification.

2. Computerized administration of perceptual speed, as carried out in the GRIP battery, was only marginally different from paper-and-pencil measures of perceptual speed. Since these measures did not offer any substantial improvements in validities over paper-and-pencil measures, the initial judgment on their usefulness would be negative.

3. Further research will be required to clarify the relationships between computerized and paper-and-pencil measures of Closure. Hidden Patterns, the best of the paper-and-pencil tests, had significant validities for Electrician's Mates, Personnelmen, and Sonar Technicians. The pattern of validities of Hidden Patterns for Sonar Technicians was duplicated by CLO-LAT, a measure which can be administered and scored automatically.

4. The two experimental tests designed to measure Movement Detection were not closely related to one another and therefore did not provide evidence of a Movement Detection factor. Instead these tests loaded on memory factors, Perceptual Speed, and perceptual Closure. On the other hand, of the measures, Memory for Patterns proved to be very useful particularly as a predictor for both specific and generalized performance of Sonar Technicians. For the Electrician's Mate and Personnelman ratings it proved to be useful at a somewhat lower level.

5. Facility in Sequential Reasoning was apparently an ability which was uniquely measurable by computer-administered tests. These tests demonstrated widespread and generalized validity for Personnelman and incremented the predictability of communication and interpersonal relations skills over that available from paper-and-pencil tests.

It is believed that the initial results with this technique are promising and that further development along these lines is warranted, particularly for jobs which require attention to scopes. Consequently, research to be carried out during Fiscal Year 1976 will be concerned with refining measures of Movement Detection, Sequential Reasoning Perceptual Closure, response latencies, and accuracy of spatial perception, together with the construction of tests for other abilities which appear to be potentially useful for personnel selection. Also, we hope to convert one or more of the tests to a branching mode designed to tailor item difficulties to candidates.

REFERENCES

French, J. W., Ekstrom, R. B., & Price L. A. *Manual for Kit of Reference Tests for cognitive factors* (Rev. 1963). Princeton: Educational Testing Service, 1963.

Gibson, J. J. (Ed.) *Motion picture testing and research: Army Air Forces aviation psychology program, report no. 7. Washington, D. C.: U.S. Government Printing Office, 1947.*

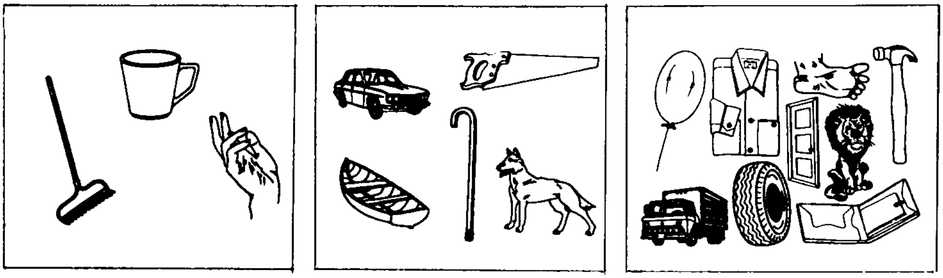
McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 347-368.

Mecham, R. C., & McCormick, E. J. *The rated attribute requirements of job elements in the Position Analysis Questionnaire.* Occupational Research Center, Purdue University, January, 1969, Report No. 1, AD-682 490. Prepared for office of Naval Research under Contract Nonr-1100(28).

APPENDIX

ILLUSTRATIVE ITEMS FROM THE EIGHT COMPUTERIZED TESTS

1. MEMORY FOR OBJECTS



2. MEMORY FOR WORDS

BIN MAN OWL

PRIZE IVORY TABLE
8TOVE MUSIC SOLID

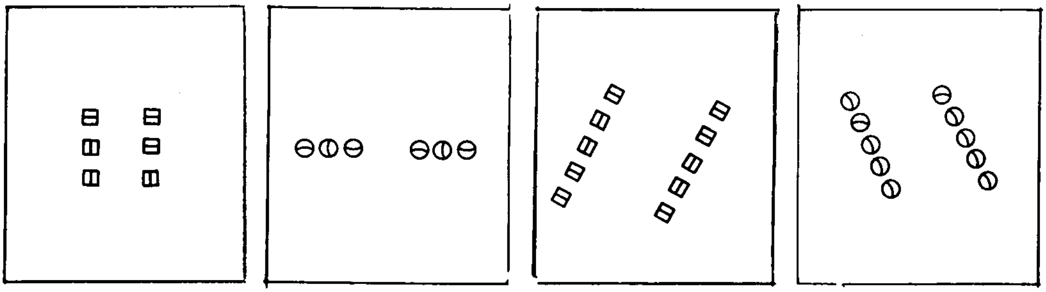
FIR TEA HAT KID ART
EYE CAT RIB BAT

3. VISUAL MEMORY FOR NUMBERS TEST

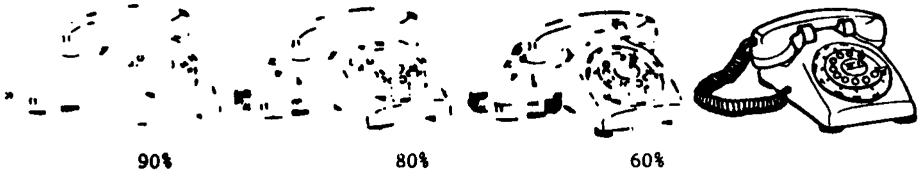
2 5 1 6*

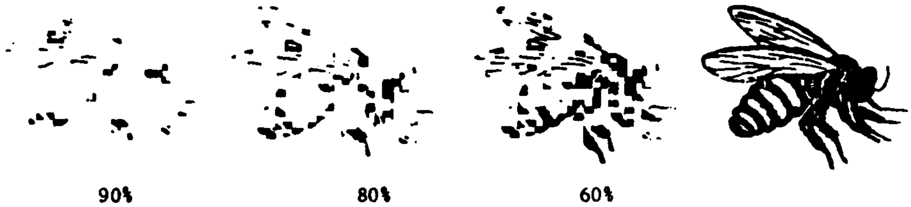
124956387*

4. COMPARING FIGURES

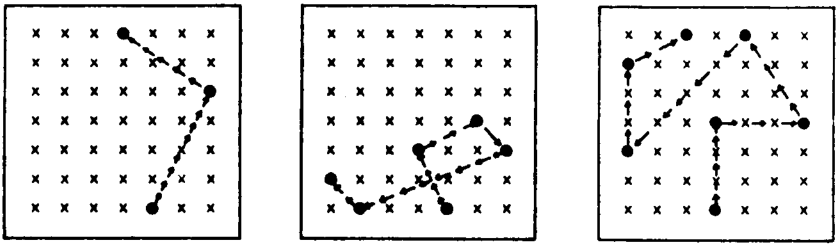


5. RECOGNIZING OBJECTS





6. MEMORY FOR PATTERNS



7. COMPUTERIZED 12 QUESTIONS

Mineral
Frequently larger than a glove

1. Is it often used as clothing?
2. Is it made of a soft material?
3. Is it often used at meals?
4. Do people often wear it?
5. Does it have moving parts?
6. Does it have a hard surface?
7. Is it always found on an auto?
8. Is it made at least partly of glass?
9. Does it have more than one use?
10. Does it use electricity?
11. Is it sometimes used by magicians?
12. Do men and women use it equally often?
13. Is it often used before a person goes out?
14. Can one use it with his eyes closed?
15. Must one touch it to use it?

16. Does it appear dark in the light?
17. Can it be used to send messages?
18. Can it improve one's appearance?

(Mirror)

8. COMPUTERIZED PASSWORD

Metal Finger	Circle	Shiny	Wedding	(Ring)
Soaring Emblem	Feathers	Large	Bald	(Eagle)

A BROAD-RANGE TAILORED TEST OF VERBAL ABILITY

FREDERIC M. LORD
Educational Testing Service

This report describes briefly a broad-range tailored test of verbal ability, appropriate at any level from fifth grade upwards, through graduate school. The test score places everyone at all levels directly on the same score scale.

In a tailored test, the items administered to an individual are chosen for their effectiveness for measuring him. Items administered later in the test are selected by computer, according to some rule based on the individual's performance on the items administered to him earlier. Improved measurement is obtained 1) by matching item difficulty to the ability level of the individual and 2) by using the more discriminating items in the available item pool. The matching of test difficulty to the individual's ability level is advantageous and desirable for psychological reasons. For references on tailored testing, see Wood (1973). Also Cliff (1975), Jensema (1974a, 1974b), Killcross (1974), Mussio (1973), Spineti and Hambleton (1975), Urry (1974a, 1974b), Waters (1974), Betz and Weiss (1974), DeWitt and Weiss (1974), Larkin and Weiss (1974), McBride and Weiss (1974), Weiss (1973, 1974), Weiss and Betz (1973).

The broad-range test consists of 182 verbal items. These were chosen from all levels of Cooperative Tests' SCAT and STEP, from the College Entrance Examination Board's Preliminary Scholastic Aptitude Test, and from the Graduate Record Examination. The choice was made solely on the basis of item type and difficulty level. There was no attempt to secure the best items by selecting on item discriminating power.

Two parallel forms of this 182-item tailored test were constructed. Only one of these forms is considered here.

Ideally there should be only one item type in each row, so that all examinees would take the same number of items of each type. The arrangement of Table 1 is an attempt to approximate this ideal using the items available. (Few if any hard items of types a and e were in the total pool; also few if any easy items of types b and c. Types a and b, also types c and e, seem fairly similar.)

TABLE 1

Broad-Range Verbal Test Items Arranged by Difficulty Level and Serial Number.
(a, b, c, d, e represent different verbal item types.)

Item Serial No.	(easy)←		Item Difficulty Level			→(hard)				
	Grade Level: IV	V	VI VII	VIII		XII				
1			a	a	a	a	a	b		
2			e	e	e	e	c			
3			d	d	d	d	d	d		
4			e	e	e	e	c	c		
5			d	d	d	d	d	d		
6			a	a	a	a	b	b		
7			e	e	e	e	c			
8		d	d	d	d	d	d			
9			e	e	e	c	c	c		
10		d	d	d	d	d	d			
11			a	a	a	a	b	b	b	b
12		e	e	e	c	c	c	c		
13			d	d	d	d	d	d		
14		e	e	e	c	c	c	c	c	
15			d	d	d	d	d	d	d	
16			a	a	a	b	b	b	b	b
17		e	e	c	c	c	c	c	c	
18	d	d	d	d	d	d	d	d		
19		e	e	c	c	c	c	c	c	
20	d	d	d	d	d	d	d	d	d	d
21		a		a	a	b	b	b	b	b
22	e	e	c	c	c	c	c	c	c	c
23		d	d	d	d	d	d	d	d	d
24	e	e	c	c	c	c	c	c	c	c
25		d	d	d	d	d	d	d	d	d

The 182 items in a single form of the test are represented in Table 1, where they are arranged in columns by difficulty level. An individual answers just one item in each row of the table—a total of just 25 items. There are five verbal item types, denoted by a, b, c, d, e. Within each item type, the items in each column are arranged in order of discriminating power with the best items at the top.

The examinee starts with an item in the first row. The difficulty level of this item is determined by the examinee's grade level, or some other rough estimate of his ability. If he answers the first item correctly, he next takes an item in the second row that is harder than (to the right of) the first item. If he answers the first item incorrectly, he next takes an item in the second row that is easier than (to the left of) the first item.

He may continue with the third and subsequent rows, moving to the right after each correct answer, or to the left after each incorrect answer, until he has at least one right answer and at least one wrong answer. At this point, the computer uses item characteristic curve theory to compute the maximum likeihood estimate of the examinee's ability level. In effect, the computer asks: For what ability level is the likelihood of the observed pattern of responses at a maximum, taking into account the difficulty and other characteristics of the items administered up to this point? The ability level that maximizes this likelihood is the current estimate of the examinee's ability.

From this point on, the next item to be administered will be of the same item type as the item in the next row that best matches in difficulty the examinee's estimated ability level. Given this item type, we survey all items of this type and administer next the item that gives the most information at his estimated ability level.

After each new response by the examinee, his ability is reestimated. The item type of the next item is determined, as above, and the best item (not already used) of that type is chosen and administered. This continues until he has answered 25 items, one for each row of the table. The maximum likelihood estimate of his ability determined from his responses to all 25 items is his final verbal ability score. According to the item characteristic curve model, all such scores, for various examinees, are automatically on the same ability scale, regardless of which set of items was administered.

About thirty different designs for a broad-range tailored test of verbal ability were tried out on the computer, administering each one to a thousand or so simulated examinees. The final design was recently chosen and has not yet been implemented on the computer for administration to real flesh-and-blood examinees.

Consider first the effect of the difficulty level of the first item administered. The vertical dimension in Figure 1 represents the standard error of measurement of obtained test score on the broad-range tailored test, computed by a Monte Carlo study. Each symbol shows how the standard error of measurement varies with ability level (horizontal axis). The four symbols represent the results obtained with four different starting points. The points marked + were obtained when the difficulty level of the first item administered was near -1.0 on the horizontal scale--about fifth grade level. The small dots represent the results when the difficulty level of the first item was near 0--about

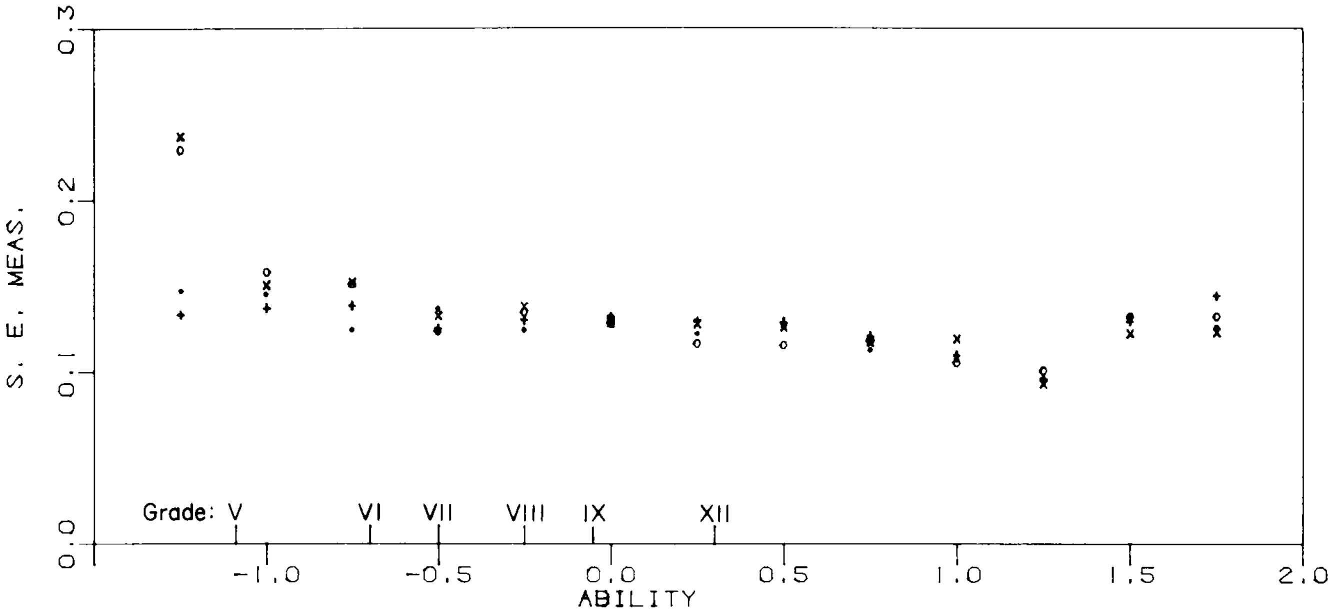


Figure 1. The standard error of measurement at 13 different ability levels for four different starting points for the 25-item broad-range tailored test.

ninth-grade level. For the hexagons, it was near 0.75--near the average verbal ability level of college applicants taking the College Entrance Examination Board's Scholastic Aptitude Test. For the points marked by an x, it was near 1.5. For any given ability level, the standard error of measurement varies surprisingly little, considering the extreme variation in starting item difficulty.

Various designs were also tried out with more columns or with fewer than the 10 columns shown in Table 1. A test with 20 columns, spanning roughly the same difficulty range as Table 1 but requiring 363 items, was found to be at least twice as good as the 10-column 182-item test of Table 1. The reason for this is not that the columns in Table 1 are too far apart, but mainly that selecting the best items (best for a particular individual) from a 363-item pool will give a much better 25-item test than selecting the same number of items from a smaller, 182-item pool. Still better tests could be produced by using still larger item pools, even though only 25 items are administered to each examinee.

It is important to compare the broad-range tailored test with a conventional test. Let us compare our broad-range tailored verbal test with the Preliminary Scholastic Aptitude Test of the College Entrance Examination Board. Figure 2 shows the information function for the Verbal score on each of three forms of the PSAT adjusted to a test length of just 25 items. Also the information function for the Verbal score on the broad-range tailored test, which administers just 25 items to each examinee. The tailored test shown in Figure 2 corresponds to the hexagons of Figure 1, since they represent the results obtained when the first item administered is at a difficulty level appropriate for average college applicants. The PSAT information functions are computed from estimated item parameters. For points spaced along the ability scale, the tailored test information function is estimated from the test responses of simulated examinees.[1]

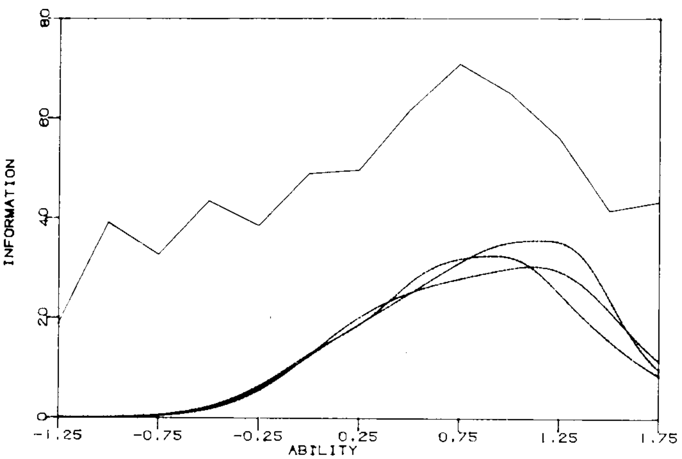


Figure 2. Information function for the 25-item tailored test, also for three forms of the Preliminary Scholastic Aptitude Test (dotted lines) adjusted to a test length of 25 items.

It is encouraging but not surprising to find that the tailored test is at least twice as good as a 25-item conventional PSAT at almost all ability levels. After all, at the same time that we are tailoring the test to fit the individual, we are taking advantage of the large item pool, using the best 25 items available within certain restrictions already mentioned concerning item type. It would, of course, be desirable to confirm this evaluation by extensive test administrations, using flesh-and-blood examinees instead of simulated examinees.

In conclusion, the writer would like to make an offer that should enable research workers and graduate students to conveniently design and build actual tailored tests and administer them to real examinees. On written request from suitably qualified individuals, he will provide estimated item parameters for the verbal items in any or all of the following Cooperative Tests:

SCAT II, Forms 1A, 2A, 2B, 3A, 3B, 4A (50 items each);

STEP II, Reading Test, Part I only, Forms 2A, 2B, 3A, 3B, 4A (30 items each);

SCAT I, Forms 2A, 2B, 3A, 3B (60 items each).

This represents a pool of 690 calibrated verbal items available for research or other purposes. (This offer expires when better methods for estimating item parameters have been developed—very soon, it is to be hoped.)

REFERENCES

Betz, N. E., & Weiss, D. J. Simulation studies of two-stage ability testing. Research Report 74-4. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Cliff, N. Complete orders from incomplete data: interactive ordering and tailored testing. *Psychological Bulletin,* 1975, *82,* 289-302.

De Witt, L. J., & Weiss, D. J. A computer software system for adaptive ability measurement. Research Report 74-1. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Jensema, C. J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology,* 1974, *27,* 29-48. (a)

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766. (b)

Killcross, M. C. A tailored testing system for selection and allocation in the British Army. A paper presented at the 18th International Congress of Applied Psychology, Montreal, August 1974.

Larkin, K. C., & Weiss, D. J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

[1] When the test score is an unbiased estimator of ability, the information function is simply the reciprocal of the squared standard error of measurement. A k -fold increase in information may be interpreted as the kind of increase that would be obtained by lengthening a conventional test k -fold.

Mc Bride, J. R., & Weiss, D. J. A word knowledge item pool for adaptive ability measurement. Reasearch Report 74-2. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

Mussio, J. J. A modification to Lord's model for tailored tests. Unpublished doctoral dissertation, University of Toronto, 1973.

Spineti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objective-based instructional programs. Unpublished manuscript, University of Massachusetts, 1975.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, *34,* 253-269. (a)

Urry, V. W. Computer assisted testing: the calibration and evaluation of the verbal ability bank. Technical Study 74-3. Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, in preparation. (b)

Waters, B. K. An empirical investigation of the stradaptive testing model for the measurement of human ability. Unpublished doctoral dissertation, The Florida State University, 1974.

Weiss, D. J. The stratified adaptive computerized ability test. Research Report 73-3. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5. Minneapolis, Minn.: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

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

Wood, R. Response-contingent testing. *Review of Educational Research,* 1973, *43,* 529-544.

SOME LIKELIHOOD FUNCTIONS FOUND IN TAILORED TESTING

FREDERIC M. LORD
Educational Testing Service

This brief note discusses some peculiar likelihood functions encountered while administering the Broad-Range Tailored Test of Verbal Ability to simulated examinees. Other workers have doubtless encountered similar problems.

Samejima (1973) shows that when the item parameters are known, there may be no finite ability level $\hat{\theta}$ that maximizes the likelihood function. Also, that the likelihood function may have more than one (local) maximum.

Barnett (1966) states "Given a single sample of observations ... [r]egularity conditions ... are no guarantee that a *single* root of the likelihood equation will exist for this sample. In fact, there will often exist multiple roots, corresponding to multiple relative maxima of the likelihood function, even if the regularity conditions are satisfied."

Huzurbazar (see Kendall & Stuart, 1973, sections 18.11-18.12) showed under regularity conditions that ultimately, as the number of observations becomes large, there is a *unique* consistent maximum likelihood estimator. His regularity conditions would apply if the test were composed of items with identical ICC. His conditions would be violated otherwise, but it should be possible to extend his proof to cover a reasonable set of regularity conditions for the present problem.

To have a large number of observations, we would need to administer a large number of test items. When the number of items is not large, and especially if the test is too hard for some individuals, we may expect $\hat{\theta}_a = -\infty$ occasionally. An examinee who makes unlucky guesses and scores below the chance level is, not unreasonably, likely to get an estimated ability of $\hat{\theta} = -\infty$. Such an estimate would presumably be corrected if a sufficiently large number of additional test items were administered to him.

In the study on a Broad-Range Tailored Test of Verbal Ability, many tens of thousands of simulated examinees took various simulated tailored tests. Items with known ICC were administered one at a time to each individual examinee. After each item was adminsitered, an approximation to the maximum likelihood estimate $\hat{\theta}$ of his ability was computed, based on all his responses up to that point.

When the examinee has wrong answers but no right answers, $\hat{\theta} = -\infty$. When he has right answers but no wrong answers, $\hat{\theta} = +\infty$. When he has both right and wrong answers, there is usually no difficulty in finding a finite $\hat{\theta}$. An occasional difficulty resolves itself as more items are administered. It is very rare to have any problem after the first ten or fifteen items, since by then the item difficulty is usually tolerably well tailored to the examinee's ability.

The present study investigates the case of simulated examinee T94 for whom there were unusual difficulties in obtaining a finite $\hat{\theta}$. Table 1 describes the first 23 items administered to him, shows his response to each item (1 = right, 0 = wrong), and gives $\hat{\theta}$, the maximum likelihood estimate of his ability based on his responses to items already administered.

Examinee T94 is really a very low ability examinee—his true θ is actually -2.9. Furthermore, the first items administered to him were very difficult items ($b_i > 1.35$) which he would have no chance at all of answering correctly except by guessing. By lucky guessing, he nevertheless got 6 items right out of the first 12.

If c_i were .20 for each of these items, the chance of a score as good or better than 6 solely by guessing is less than .02. The maximum likelihood estimates of the examinee's ability based on his performance on these first twelve items range from 1.6 to 2.2, as shown in the last column of the table.

His guessing on the next seven items was uniformly unsuccessful. All items through item 17 were difficult, with $b_i > 1.35$. His performance on these 17 difficult items earned him an ability estimate of $\hat{\theta} = 1.2$.

Item 18 was an easier item, $b_{18} = .65$. I suggest that the following rationalizations provide a correct explanation of the θ subsequently obtained.

The examinee has answered correctly 6 items with $b_i > 1.35$ and has failed 12 items including one with $b_i = .65$. The last failure suggests that θ is low and that earlier correct responses were due to lucky guessing. If θ is low, all items so far administered are too difficult for the examinee and are of no use, even for placing a lower bound on his ability level. When an examinee has given only wrong responses and lucky random guesses, his estimated ability should be $\hat{\theta} = -\infty$.

When the examinee answers item 20 ($b_{20} = -.83$) correctly, it is now plausible to assume that his ability lies between $-.83$ and .65 (.65 being the difficulty level of item 18, which he answered incorrectly). The maximum likelihood estimate turns out to be $\hat{\theta} = -.4$.

Research reported in this paper has been supported by grant GB-41999 from National Science Foundation.

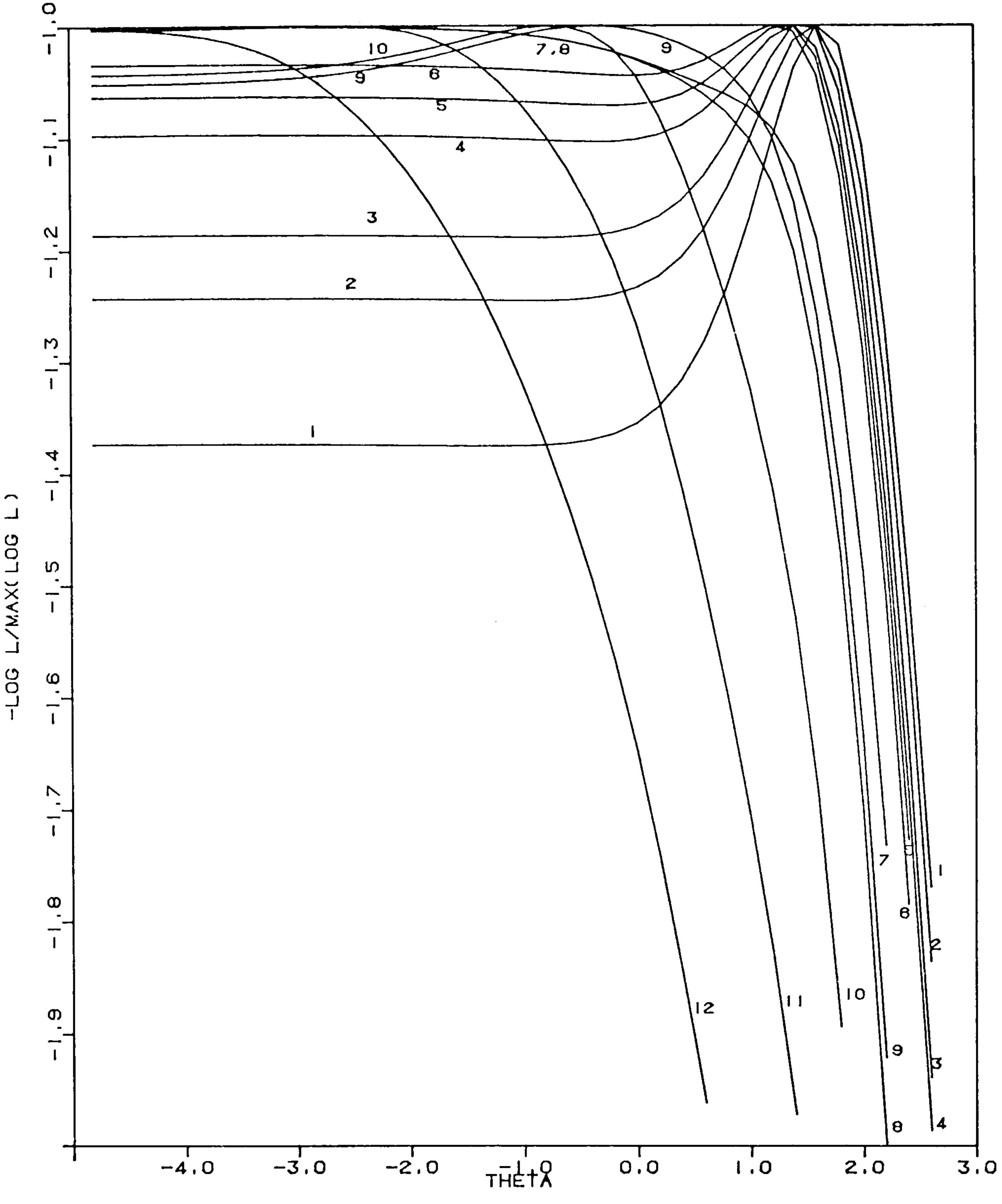


Figure 1. Standardized likelihood function for examinee no. T94, $\theta = -2.9$.

TABLE 1

Successive Estimates of Ability for Examinee T94

Item no.	Curve no. in Fig. 1	Item Parameters a	b	c	Examinee's response	Number of right answers	Log likelihood* at $\hat{\theta}$	Estimated ability** $\hat{\theta}$
1		.61	2.20	.19	0	0		–
2		2.05	1.74	.18	1	1		2.2
3		1.48	2.51	.17	0	1		1.9
4		1.89	1.96	.20	0	1		1.6
5		1.93	1.89	.24	1	2		1.8
6		2.21	1.73	.21	1	3		2.0
7		1.57	1.76	.07	0	3		1.8
8		1.68	1.40	.15	1	4		1.8
9		1.42	1.36	.13	0	4		1.7
10		1.27	1.65	.28	1	5		1.7
11		1.56	1.49	.19	0	5		1.6
12	1	1.34	1.54	.19	1	6	–7.7	1.6
13	2	1.07	1.52	.20	0	6	–8.7	1.6
14	3	1.31	1.89	.09	0	6	–9.2	1.4
15	4	.93	1.35	.20	0	6	–10.1	1.4
16	5	1.02	1.98	.21	0	6	–10.7	1.4
17	6	1.03	1.88	.13	0	6	–11.1	1.2
18	7	1.24	.65	.20	0	6	–11.7	– ∞
19	8	2.00	1.27	.10	0	6	–11.8	– ∞
20	9	.88	–.83	.33	1	7	–12.3	–.4
21	10	2.10	.05	.21	0	7	–12.6	–.8
22	11	1.37	–1.49	.15	0	7	–13.3	-2.6
23	12	1.10	–2.84	.24	0	7	–13.6	– ∞

*Not computed for $n < 12$.

**For $n = 2,3,\ldots,11$, the listed $\hat{\theta}$ is an approximate value determined numerically. For $n > 11$, the listed was $\hat{\theta}$ was read from values of the log likelihood tabulated at intervals of .2 along the θ scale.

Subsequent failures on items 21 and 22 lower this estimate to -.8 and then to -2.6. When the examinee finally fails an item with $b_i = -2.84$, it now appears that all earlier correct answers were due to lucky guessing and that all items so far administered were too difficult for this examinee. The situation is much the same as the situation after the answer to item 18, already discussed. Again, not unreasonably, $\hat{\theta} = -\infty$.

In this testing, only the very last item was of appropriate difficulty for the examinee, whose true ability was $\theta = -2.9$. All but the last two items were very much too hard. He answered both the last two items incorrectly. Thus, it is only to be expected that his final ability estimate is $\hat{\theta} = -\infty$. Administration of further items of appropriate difficulty would quickly correct this estimate.

The likelihood functions used to obtain most of the successive $\hat{\theta}$ discussed above are shown in Figure 1. The code numbers identifying the curves are given in Table 1. In order to get them all on the same graph, each likelihood function is divided by its maximum value, so that the maxima of the normalized curves all fall on the top boundary of the figure. These curves, together with the discussion given above, seem to explain the anomalous values of $\hat{\theta}$. When enough responses have been obtained to indicate a lower limit to the examinee's ability, then finite ability estimates will be obtained.

REFERENCES

Barnett, V. D. Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, 1966, *53*, 151-165.

Kendall, M. G. and Stuart, A. *The advanced theory of statistics.* New York: Hafner, Vol. 1, 1969; Vol. 2, 1973.

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

BAYESIAN TAILORED TESTING AND THE INFLUENCE OF ITEM BANK CHARACTERISTICS [1]

CARL J. JENSEMA
Gallaudet College

Conventional tests are generally constructed to discriminate over a rather wide range of examinee ability. One of the consequences of this approach is that a conventional test usually contains many items which are not appropriate for a particular level of ability. Psychometricians have long been aware of this and in recent years they have increasingly turned their attention to the possibility of programming computers to design and administer tests.

Of the many computerized testing methods which have been proposed, the Bayesian process developed by Owen (1969) seems to be the most elegant and intuitively appealing method. It assumes locally independent binarily scored items and a normal ogive model (Lord and Novick, 1968, Ch. 16) in which the probability of passing a free response item g at ability level θ is expressed as

$$P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} \exp\left[\frac{-t^2}{2}\right] dt \tag{1}$$

If the item is not of the free response type and c_g is the probability of guessing correctly, the probability of passing becomes

$$P_g'(\theta) = P_g(\theta) + c_g\,[1 - P_g(\theta)] \tag{2}$$

The derivation of Owen's Bayesian tailoring process has been described several times in the literature (Owen, 1969; Urry, 1971; Jensema, 1974a). We will briefly run through the fundamental formulas here for the sake of completeness.

Suppose $N(\theta_O, \sigma_O{}^2)$ expresses our knowledge of an examinee having ability θ. If we administer free response item g, which has discrimination and difficulty parameters a and b, and if the examinee responds correctly, Bayes' theorem specifies that the information available is

$$P(\theta|1) = k\,P_g(\theta)\,(\sqrt{2\pi}\,\sigma_o)^{-1} \exp\left[(\frac{-(\theta - \theta_o)^2}{2\sigma_o{}^2}\right] \tag{3}$$

where $P_g(\theta)$ is defined by (1) and k is such that

$$\int_{-\infty}^{\infty} P(\theta|1)\,d\theta = 1. \tag{4}$$

The solution is

$$k^{-1} = 1/2\,(1 - \operatorname{erf} D) \tag{5}$$

where erf D is the error function

$$\operatorname{erf} D = \frac{2}{\sqrt{\pi}} \int_0^D \exp(-t^2)\,dt \tag{6}$$

and

$$D = \frac{b - \theta_o}{\sqrt{2\,(a^{-2} + \sigma_o{}^2)}} \quad . \tag{7}$$

The expectation of the posterior mean is

$$E(\theta|1) = \theta_o + \frac{\sqrt{2}\,\sigma_o{}^2}{\sqrt{\pi\,(a^{-2} + \sigma_o{}^2)}} \exp(-D^2)\,(1 - \operatorname{erf} D)^{-1} \tag{8}$$

and the variance is

$$\operatorname{var}(\theta|1) = \sigma_o^2 \left[1 - \frac{\frac{2}{\sqrt{\pi}} - 2D\exp(D^2)\,(1 - \operatorname{erf} D)}{\sqrt{\pi}(1 + a^{-2}\sigma_o{}^{-2})\,(\exp(D^2)\,(1 - \operatorname{erf} D))^2}\right] \tag{9}$$

Similarly, if the examinee gives a wrong response to item g we have

$$P(\theta|O) = \frac{k}{k-1}(1 - P_g(\theta))\,(\sqrt{2\pi}\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2\sigma_o{}^2}\right], \tag{10}$$

$$E(\theta|O) = \theta_o - \frac{\sqrt{2\sigma_o{}^2}}{\sqrt{\pi\,(a^{-2} + \sigma_o{}^2)}} \exp(-D^2)\,(1 + \operatorname{erf} D)^{-1}, \tag{11}$$

[1] This research was supported by the Office of Demographic Studies, Gallaudet College, Washington, D. C., 20002.

and

$$\text{var}\,(\theta\,|O) = \sigma_o^{\;2}\left[1 - \frac{\frac{2}{\sqrt{\pi}} + 2D\exp(D^2)\;(1+\operatorname{erf} D)}{\sqrt{\pi}\,(1+a^{-2}\sigma_o^{\;-2})\;(\exp(D^2)\;(1+\operatorname{erf} D))^2}\right] \tag{12}$$

To expand this discussion a little further assume that item g is not a free response item and that it has a probability C_g of guessing correctly. If the examinee gives a correct response we have

$$P'(\theta|1) = \lambda P_g'(\theta)\;(\sqrt{2\pi}\,\sigma_o)^{-1}\exp\left[\frac{-(\theta-\theta_o)^2}{2\,\sigma_o^{\;2}}\right], \tag{13}$$

$$E'(\theta|1) = \theta_o + (1 - C_g)\,k^{-1}\,\lambda S\,, \tag{14}$$

and

$$\text{var}'(\theta|1) = \sigma_o^{\;2} - (1-C_g)\,k^{-1}\,\lambda S^2\;(t - C_g\,\lambda) \tag{15}$$

where the prime is used to signify the effect of guessing, $P_g'(\theta)$ is defined by (2), and we take

$$\lambda^{-1} = C_g + (1-C_g)\,k^{-1}\,, \tag{16}$$

$$S = k\,\sigma_o\exp(-D^2)\;(2\pi(1+a^{-2}\sigma_o^{\;-2}))^{-\frac{1}{2}} \tag{17}$$

$$\text{t} = 1 - 2\sqrt{\pi}\,\text{k}^{-1}\,D\exp(D^2). \tag{18}$$

If the examinee gives a wrong response the formulas in (10), (11), and (12) hold, since our information, that the examinee does not know the correct answer, is the same as in the free response case.

Now assume we have n items and want to select the best one for administration. The expected posterior variance of θ after administration of a particular item is

$$\underset{u}{E}(\text{var}\,(\theta|u) = \theta_o^{\;2} + \sigma_o^2 - P(O)\;[E(\theta|O)]^2 - P(1)\;[E(\theta|1)]^2$$

$$= \sigma_o^{\;2}\left[1 - \frac{2}{\pi\,(1+a^{-2}\sigma_o^{\;-2})\exp(2D^2)\;(1-(\operatorname{erf} D)^2)}\right] \tag{19}$$

when items are of the free response type and

$$\underset{u}{E'}(\text{var}\,(\theta|u)) = \sigma_o^{\;2}\left[1 - \frac{(1-C_g)\,\lambda\,(1+C_g\,(1-k^{-1}))}{2\pi\,(1+\sigma_o^{\;-2}a^{-2})\;(1-k^{-1})\exp(2D^2)}\right] \tag{20}$$

when the items are affected by guessing. In (19) and (20) u refers to the correctness of the examinee's response and is taken as 1 or 0. The item which leads to the smallest expected posterior variance is the most desirable one to administer. It is sufficient to select the item with the smallest value α where

$$\alpha = (a^{-2} + \sigma_o^{\;2})\;(1 - (\operatorname{erf} D)^2)\exp(2D^2) \tag{21}$$

for free response items and

$$\alpha' = (\frac{1}{1 - C_g})\;(1+\sigma_o^{\;-2}a^{-2})\;(1-k^{-1})\,\lambda^{-1}\exp(2D^2). \tag{22}$$

when guessing is present.

If we have a pool of n items and estimates of the normal ogive model parameters for each item, we may use a Bayesian sequential procedure to select items for administration to a particular examinee. Let $\hat{\theta}_{(m)}$ and $\hat{\sigma}^2_{(m)}$ be an estimate of the examinee's ability and its variance where m indicates the number of items administered. Assume the population has ability distributed as $N(0,1)$ and take $\hat{\theta}_{(o)}$ and $\hat{\sigma}^2_{(o)}$ as 0 and 1. Calculate α_i values for all (unused) items, $i = 1,2,\ldots..,$ (n–m), using (22). (We will assume that the items are not free-response.) The examinee is administered the item with the smallest α_i value. If an incorrect response is given, $\hat{\theta}_{(m+1)}$ and $\hat{\sigma}^2_{(m+1)}$ are calculated from (11) and (12). If the response is correct, (14) and (15) are used. This cycle is repeated until $\hat{\sigma}_{(m)}$ is within some pre-selected limit. The selection of a $\hat{\sigma}_{(m)}$ value for termination is, of course, arbitrary. It is usually selected to yield some expected level of validity according to

$$r_{\theta\hat{\theta}} = \sqrt{1 - \hat{\sigma}^2_{(m)}} \tag{23}$$

The characteristics of an item bank used for tailored testing are very important to the efficiency and accuracy of the process. There are four basic requirements for a good item bank. These have been mentioned in whole or part in a number of publications (i.e. Urry, 1970, 1971, 1971b, 1974; Jensema, 1972, 1974a, 1974b; etc.) and may be summarized as follows:

1) Item discrimination should be as high as possible and should not be less than .8.
2) Item guessing probabilities should be as low as possible.
3) The item bank must consist of a sufficiently large number of items.
4) Item difficulties should have a rectangular distribution.

The remainder of this paper will concentrate on demonstrating the importance of each of these four requirements.

Assume that an infinitely large item bank exists and that all items have the same discriminatory power and the same probability of guessing correctly. The assumption of an infinitely large item bank allows the selection of an item *i* having a difficulty level exactly equal to any given estimate of ability. When this can be done many of the formulas may be greatly simplified since we have:

$$D_i = 0 \tag{24}$$

and

$$\text{erf}\, D_i = 0. \tag{25}$$

The equations for $\hat{\sigma}^2_{(m\ +\ 1)}$ for correct and incorrect responses become

$$\hat{\sigma}^2_{(m\ +\ 1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i\,(1 - C_i)^2}{\pi\,(1 - C_i)^2} \right] \tag{26}$$

and

$$\hat{\sigma}^2_{(m\ +\ 1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i}{\pi} \right] \tag{27}$$

where *m* is the number of items previously administered.

An item *i*'s difficulty is the point at which the probability of knowing the correct answer is exactly .5. If guessing is in effect the probability of responding correctly is equal to the probability of knowing the answer plus the probability of guessing correctly. Then $\hat{\sigma}^2_{(m+1)}$ may be expected to be the sum of (26) and (27) weighted by the probabilities of a correct or incorrect response:

$$E\,\hat{\sigma}^2_{(m\ +\ 1)} = \hat{\sigma}^2_{(m)} \left[.5\,(1 + C_i) \left(1 - \frac{2t_i\,(1 - C_i)^2}{\pi\,(1 + C_i)^2} \right) + .5\,(1 - C_i) \left(1 - \frac{2t_i}{\pi} \right) \right] \tag{28}$$

A little algebraic manipulation reduces this to

$$E\,\hat{\sigma}^2_{(m\ +\ 1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i\,(1 - C_i)}{\pi\,(1 + C_i)} \right] \tag{29}$$

Inserting appropriate values for a_i and c_i in equation (29) and plotting the results against the number of items administered demonstrates the influence of item discrimination and guessing probability on the tailoring process. Figure 1 plots the expected standard error of the estimate $\epsilon\hat{\sigma}_{(m+1)}$ by the number of items administered for five levels of discrimination when guessing probability is zero and an infinite number of items are available. Notice the sharp difference in the number of items needed at different levels of discrimination. For example, if the items have discriminatory powers of 2.5 only 4 or 5 items are needed to reach a standard error of the estimate of .30 while 17 or 18 items are needed to reach this level when item discrimination is only 1.0.

Now suppose we take item discrimination to be 1.0, a rather low value which is easily obtained. Figure 2 plots the expected standard error of the estimate for various guessing values by the number of items administered. The guessing values range from .5 (i.e. true-false items) to 0.0 (i.e. free response items.) The greater the probability of guessing, the more items required to reach a specific standard error of the estimate.

To give a clear example of the combined effects of discrimination and guessing on the tailoring process, suppose we have three item banks which, for convenience, are referred to as I, II, and III. Assume Bank I items have discrimination and guessing paramenters of .5 and .33. Bank II's parameters are 1.0 and .25 while Bank III has parameter values of 2.0 and .20. These banks may be roughly classified as *unacceptable, fair,* and *excellent* for tailored testing purposes. Assuming that each bank has an infinite number of items and plotting the expected standard error of the estimate against the number of items administered, the three curves in Figure 3 are obtained.

In Figure 3, notice that Bank I would give unacceptable results. After 30 items the expected standard error of the

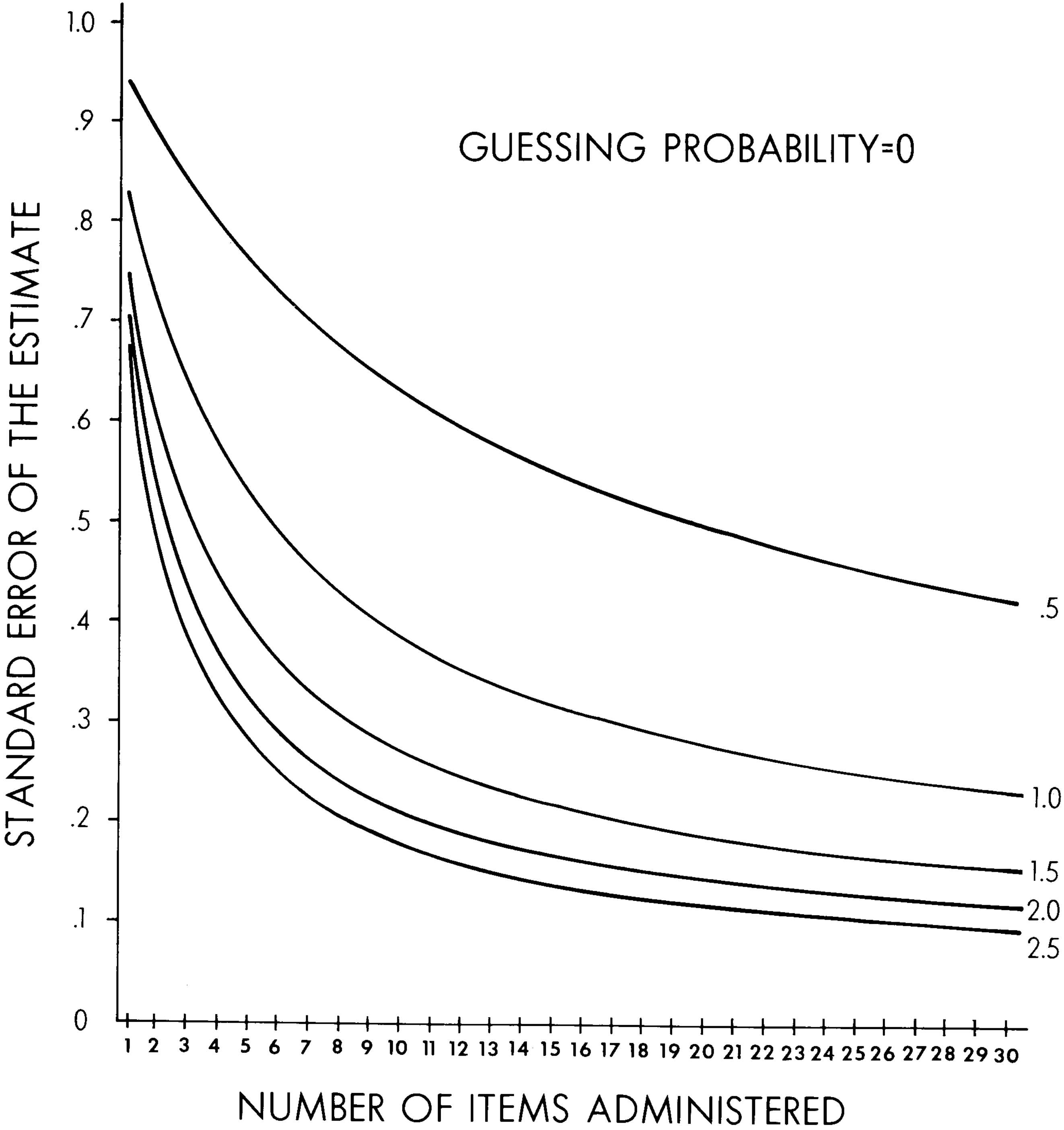


Figure 1. Expected standard error of the estimate according to number of items administered at five levels of item discrimination.

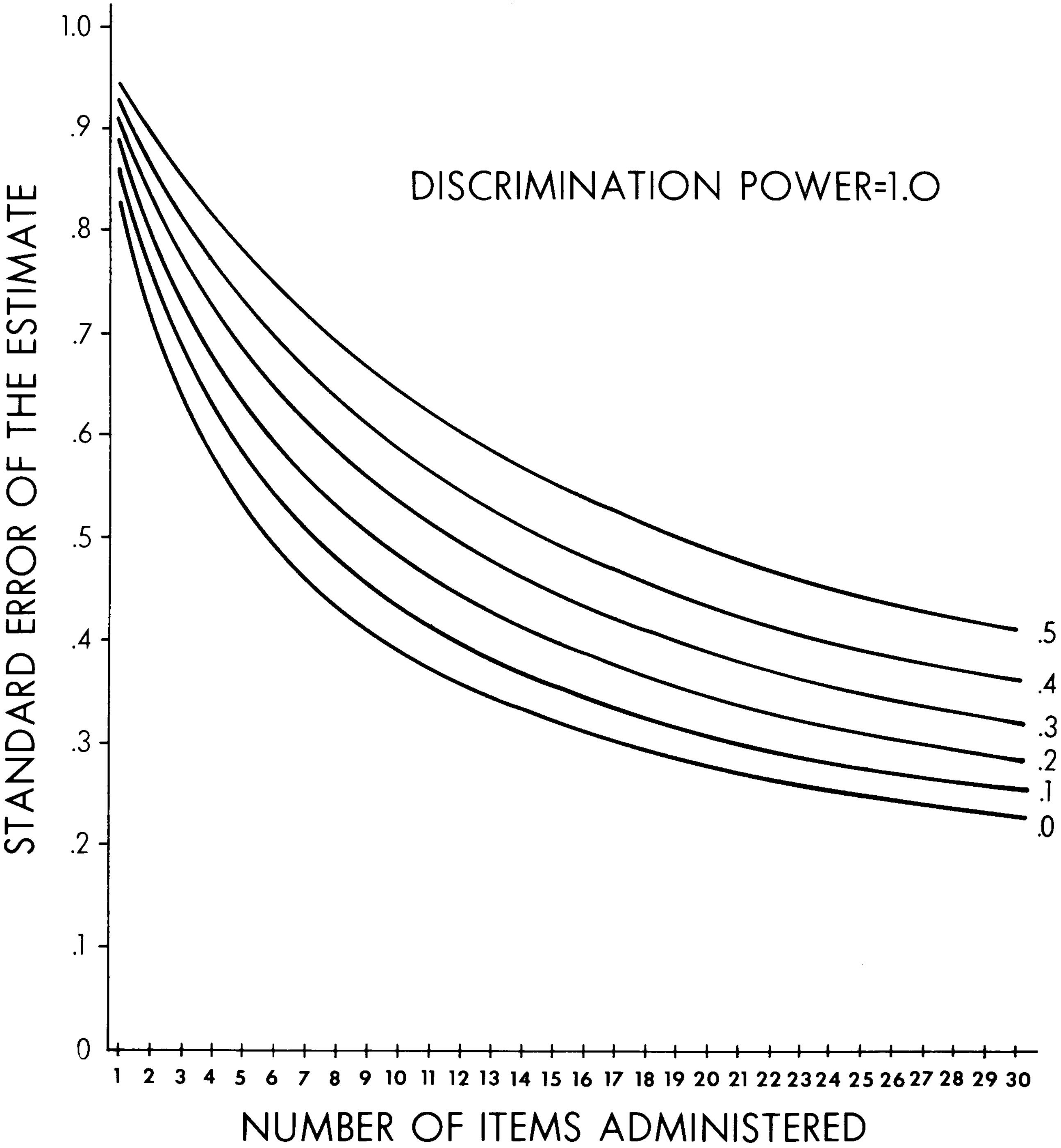


Figure 2. Expected standard error of the estimate according to number of items administered at six guessing probabilities.

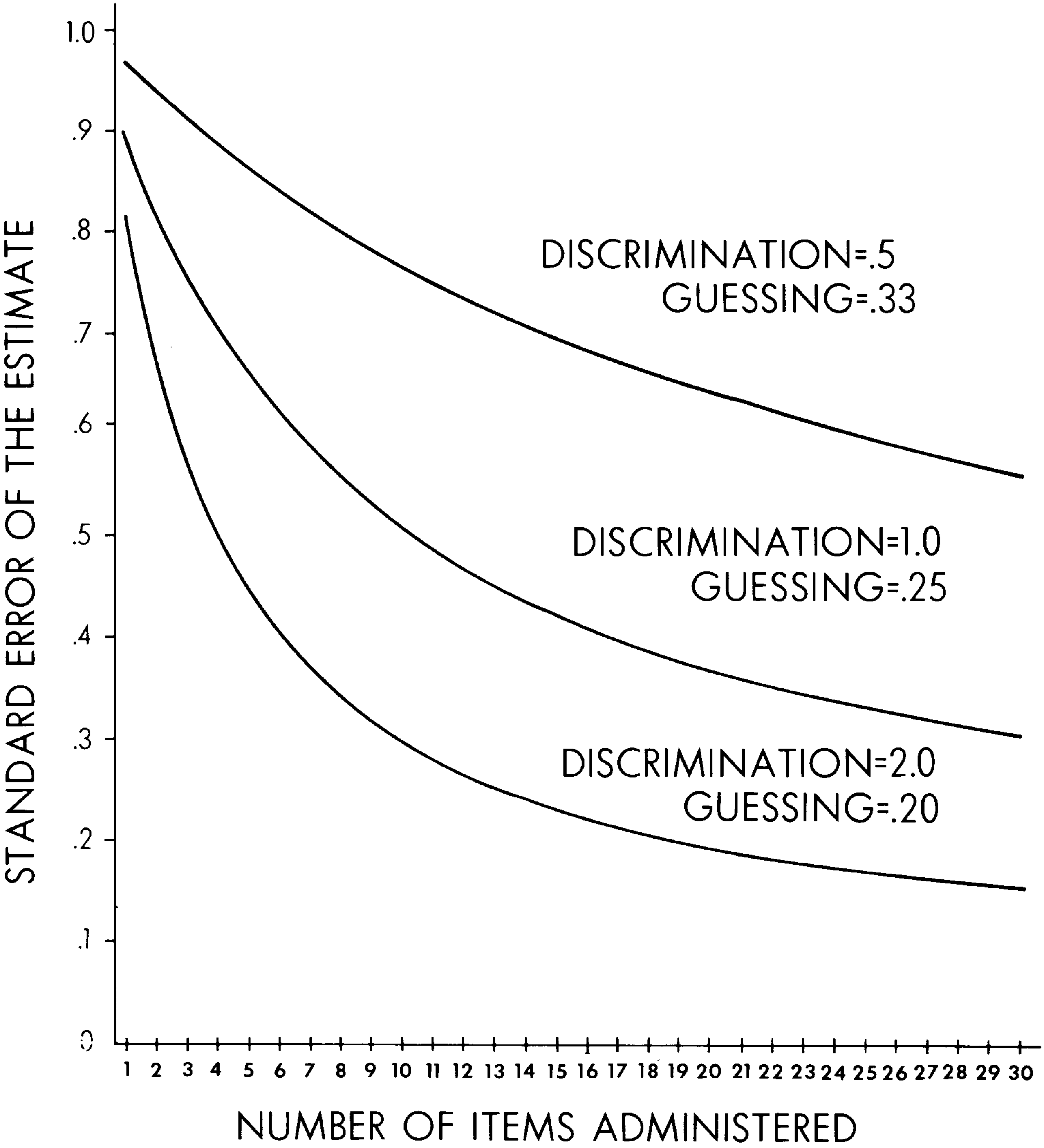


Figure 3. Expected standard error of the estimate for three item banks according to number of items administered.

estimate is only .56 (i.e. reliability = .69, validity = .83). In contrast an excellent item bank, such as Bank III, would reach this level after only 3 or 4 items. The advantage of high discrimination and low guessing probability in an item bank is obvious.

Up to this point we have discussed the behavior of Bayesian tailored testing when the item bank is assumed to be of unlimited size. The obvious question which follows is what happens when item bank sizes are within practical limits? To answer this question, Monte-Carlo data for 200 items are generated for each of 100 "examinees" using Urry's (1970) "LOGIST" program. The parameters for discrimination (1.0) and guessing (.25) were the same as for Bank II mentioned earlier. Eight sets of 25 difficulty values (-2.4, -2.2, . . . , 0.0, . . . , 2.2, 2.4) were employed. Bayesian tailored testing was simulated with this data using 50, 75, 100, 150, and 200 items in the bank. Since difficulty had been specified in sets of 25 values, the item banks had 2, 3, 4, 6, and 8 items at each of the 25 difficulty levels respectively.

For each of the five item banks and for each of the 100 examinees, tailoring was simulated until 30 items had been "administered". As each item was "administered" the new estimate of ability was recorded. Since the data was randomly generated, true ability (distributed as $N(0,1)$ was known and could be correlated with estimated ability. Table I gives the validity (correlation between true and estimated ability) for each item bank by the number of items "administered". The last column in Table I gives the expected validities for an item bank of infinite size as calculated from equation (32) and (23).

The Monte-Carlo data above represents items which are passable but not especially good for tailored testing. To see how item bank size would influence validity when the bank was composed of excellent items, the Monte-Carlo data tailoring simulation was repeated with higher discrimination

TABLE 1

Validity ($r_{\theta\hat{\theta}}$) Obtained With Different Size Item Banks
(Monte-Carlo Data, N=100, A=1.0, C=.25)

	ITEMS IN BANK					
Items Administered	50	75	100	150	200	∞*
1	.53	.53	.53	.53	.53	.44
2	.59	.59	.59	.59	.59	.57
3	.65	.65	.65	.65	.65	.66
4	.72	.72	.72	.72	.72	.72
5	.78	.78	.78	.78	.78	.76
6	.81	.80	.80	.80	.80	.79
7	.83	.82	.82	.82	.82	.81
8	.84	.84	.84	.84	.84	.83
9	.85	.85	.84	.84	.84	.85
10	.86	.86	.86	.85	.85	.86
11	.86	.87	.88	.87	.87	.87
12	.87	.87	.89	.87	.87	.88
13	.89	.89	.89	.87	.88	.89
14	.90	.91	.90	.88	.88	.90
15	.91	.91	.91	.90	.90	.91
16	.91	.92	.92	.91	.91	.91
17	.92	.92	.92	.92	.91	.92
18	.92	.92	.93	.92	.92	.92
19	.92	.92	.93	.92	.92	.93
20	.93	.93	.93	.93	.93	.93
21	.93	.93	.93	.93	.93	.93
22	.93	.94	.94	.94	.93	.94
23	.93	.94	.94	.94	.94	.94
24	.93	.94	.94	.94	.94	.94
25	.93	.94	.95	.94	.94	.94
26	.94	.95	.95	.94	.94	.95
27	.94	.95	.95	.94	.95	.95
28	.94	.95	.95	.95	.95	.95
29	.94	.95	.95	.95	.95	.95
30	.94	.95	.95	.95	.95	.95

*Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items.

TABLE 2

Validity ($r_{\theta\hat{\theta}}$) Obtained With Different Item Bank Sizes
(Monte-Carlo Data, N=100, A=2.0, C=.2)

	ITEMS IN BANK					
Items Administered	50	75	100	150	200	∞*
1	.66	.66	.66	.66	.66	.58
2	.75	.75	.75	.75	.75	.74
3	.84	.84	.84	.84	.84	.82
4	.89	.89	.89	.89	.89	.86
5	.92	.92	.92	.92	.92	.90
6	.93	.93	.93	.93	.93	.91
7	.94	.94	.94	.94	.94	.93
8	.95	.95	.95	.95	.95	.94
9	.96	.95	.95	.95	.95	.95
10	.96	.96	.96	.96	.96	.96
11	.97	.96	.96	.96	.96	.96
12	.97	.96	.96	.96	.97	.96
13	.97	.97	.97	.97	.97	.97
14	.97	.97	.97	.97	.97	.97
15	.97	.97	.98	.97	.98	.97
16	.97	.98	.98	.98	.98	.98
17	.97	.98	.98	.98	.98	.98
18	.98	.98	.98	.98	.98	.98
19	.98	.98	.98	.98	.98	.98
20	.98	.98	.98	.98	.98	.98
21	.98	.98	.98	.98	.98	.98
22	.98	.98	.99	.98	.98	.98
23	.98	.98	.99	.98	.98	.98
24	.98	.98	.99	.98	.98	.98
25	.98	.98	.99	.99	.99	.98
26	.98	.98	.99	.99	.99	.99
27	.98	.98	.99	.99	.99	.99
28	.98	.98	.99	.99	.99	.99
29	.98	.98	.99	.99	.99	.99
30	.98	.98	.99	.99	.99	.99

*Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items.

(2.0) and lower guessing (.20) parameter values. These configurations correspond to Bank III mentioned earlier. The results of the simulated tailoring with this new data are given in Table 2.

For practical application it is apparent that a very large number of items is not a critical item bank characteristic if the bank is good in other respects. In both Table 1 and Table 2 the Monte-Carlo data validities obtained for the five banks closely match each other and they also parallel the validities to be expected from a corresponding item bank of infinite size. However, it must be remembered that this was Monte-Carlo data and the tailoring simulation used known parameter values for discrimination, difficulty, and guessing. With real data involving imprecise parameter estimates and a possible non-uniform distribution of difficulty, it would be wise to be a bit cautious if a bank had, say, fewer than 75 items. In connection with this, there are some practical problems which arise if an item bank is too large. A large bank has more items available for administration, but the storage requirements and the increased computer processing needed for item selection also slow things down while adding to overall computer costs. (Some good cost-efficiency studies are needed on this!)

The last item bank requirement is uniform distribution of difficulty. The exact results of violating this rule are difficult to predict, since they would necessarily depend on the actual distribution of item difficulty, the discrimination and guessing parameter values, the number of items in the bank, and the criteria used to terminate the tailoring process. The essential point to remember is that the Bayesian tailoring procedure attempts to select for administration the item which will yield the most information. If, at a particular level of difficulty, there are no items available, the Bayesian process will be forced to select an item which is not appropriate and which will yield less than an optimal amount of information.

To summarize, this paper has outlined a Bayesian approach to item selection for tailored testing. Four basic requirements of a good item bank for this process have been discussed. If these requirements are met, Bayesian tailored testing will yield excellent results. The key to the process lies in careful construction of item banks. If attention is given to this, the Bayesian tailoring process gives us a fundamental tool for practical application of latent trait mental test theory.

REFERENCES

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C. J. An application of latent trait mental test theory. *Br. J. Math. Statist. Psychol.* 27, 29-48, 1974a.

Jensema, C. J. The validity of Bayesian tailored testing. *Ed. and Psychol. Meas.* 34, 757-766, 1974b.

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

Owen, R. J. A Bayesian approach to tailored testing. *Res. Bull.* 69-92. Princeton, N. J.: Educational Testing Service, 1969.

Urry, V. W. A Monte-Carlo investigation of logistic mental test models. Unpublished doctoral dissertation. Purdue University, 1970.

Urry, V. W. Approximation methods for the item parameters of mental test models. *Res. Bull.* 0871-202. Seattle: University of Washington, Bureau of Testing, 1971a.

Urry, V. W. Individualized testing by Bayesian estimation. *Res. Bull.* 0171-177. Seattle: University of Washington, Bureau of Testing. 1971b.

Urry, V. W. Approximations of item parameters of mental test models and their uses. *Ed. and Psychol. Meas.* 34, 353-369, 1974.

REFLECTIONS ON ADAPTIVE TESTING

DUNCAN N. HANSEN
Memphis State University

The purpose of this paper will be to reflect on various aspects of the adaptive testing field. Building from our prior Memphis State University and Air Force work in the area, the various issues, alternatives, priorities and ultimate styles of research for adaptive testing will be placed in the context of empirical findings and institutional requirements. The rationale for proposing such a pontifical and extremely challenging task is twofold. First, all our substantive and empirical work was recently reported (Hansen, 1975) and it would seem superfluous to rewrite or try to extend this research prior to more effort; therefore, only the major questions and findings will be summarized in this paper. Secondly, the various characteristics of the adaptive testing field will be reflected on in terms of research productivity and institutional requirements. Having by scholarly necessity been forced to read extensively in this domain over the past five years and, in many instances, to take a pencil in hand to follow a variety of formal derivations, I think it appropriate for me to comment about various purposes and styles of research. This is not done to criticize any of these models but rather to seriously address the question, "Are we moving in the most profitable direction and using the most expeditious procedures?"

MSU Adaptive Testing

Generic to any research in adaptive testing or that relating to the whole educational enterprise is a clear understanding of its purpose. For our group, the purpose is that of facilitating achievement or mastery testing. Within industry and military training it is common to find that testing time and managerial demands, especially for individualized techniques, are now taking upwards of 20 percent of the total training time. Such a training commitment becomes sizable and the systems managers must inevitably ask the question, "Is there a more efficient and effective way of going about it?" For example, the Air Force Advanced Instructional System will ultimately have 700 students aboard for any given training shift (2,100 students per day). If one considers that their day consists of six hours of instruction and that approximately 20 percent of this will be given over to testing, one can see that 72 minutes are being allocated on the average for each student's evalution per day. If such testing time can be reduced by 50 percent, an adaptive testing goal set for our efforts, then effectively 1.5+ million dollars worth of salaried money can be gained by shortening the training time for the 2,100 manpower units flowing in this system. It is precisely this type of monetary achievement that impresses our representatives in Congress concerning the importance of research ideas applied to significant educational problems. As will be suggested later, such specific, operational goals, while unachieved to date, give the best rationale for continued research support in this area.

As a corollary to the efficiency issue, an accompanying objective concerns the efficient application of computer technology to the testing process. In essence, one can demonstrate that adaptive testing falls closer to the drill and practice end of the computer usage continuum (Hansen, et. al., 1973) and certainly is orders of magnitude less demanding on a computer than CAI or simulated training. Our experiences and computer algorithms can be offered to you for your consideration. These document an efficient use of computers, tools which are fast becoming integral to the educational processes within our human institutions.

Finally, adaptive testing should be considered within the context of a total systems effort. For our group, adaptive testing is just one component within an overall adaptive instructional system. As one significantly alters the environment and the sequence of educational elements so as to foster or optimize learning outcomes for a given individual, one can see that testing becomes just one more component in such a stream of events. One should look at it, though, in terms of its contributions to the individual and the institution, be this increasing levels of competency or the educational system itself. Thus, one can contend that theoretical models have little or no value unless placed within such a system context since it is the context which will mold and determine the criteria, values, and operation by which its characteristics shall be judged. Let us turn then, to the specifics of the MSU adaptive testing model.

MSU Adaptive Testing Model

Our adaptive testing approach involves three components, namely, the entry of a student into the test, tailoring the test items for the student, and adaptive scoring procedure. Each of these will be discussed in turn. In reference to the entry and test composition processes, a student is entered at a level commensurate with our prediction of his ultimate performance. Therefore, using linear regression techniques mostly composed of variables from prior test performances, a student is placed into a monotonically arranged test. Such a procedure seems to work quite successfully and has an additional advantage of reducing the number of test items to be presented for any given student. (How this is done should be obvious given an understanding of the flexilevel algorithm.)

While we have very limited data concerning the efficacy of this procedure, entry to final score correlations tend to be in the low .80 range. These are similar to correlation coefficients reported by Cleary at the University of Wisconsin for students who were placed in a branch test according to a predicted outcome level (a personal communication at an AERA conference in 1969). Thus, the adaptive entry of a student seems to be a positive step forward and should be taken into account by any model working within this field.

In reference to test composition, it can be specified that each student, based on his entry profile, will have a specially developed set of composed items. These composed items may reflect information concerning the student's prior performance on various objectives which form the achievement test. Therefore, if one has information about a student's achievement of these objectives, there is no rationale for presenting the item. It is precisely this concept of test composition that appears so advantageous, although it has not been empirically pursued. One can anticipate that sometime within the next year one of the military training systems will pursue it in greater depth.

Tailored Testing of Items

As indicated, Lord's flexilevel algorithm is utilized for tailoring the presentation of test items. For achievement testing, this approach violates the assumptions as to normality as axiomatically represented within this model, but it can empirically be countered that our findings justify the utilization of the algorithm from a student and systems point of view. This adaptation is precisely the ability to move between very difficult and very easy items while at the same time adjusting cutoff criteria where considered appropriate (up to this point our group always used end of test item cutoff procedures but others could be considered). Achievement and mastery testing, especially in a technical training environment, always tend to yield asymmetric performance score distributions. Such distributions, if better understood, could be more readily adapted to flexilevel testing and yield optimal algorithms. Obviously, no attempt to prove such an assertion has been made at this point.

Scoring

Our views on scoring represent an attempt to remain consistent with the traditional procedures of adding up all correct responses and giving weights to those items that are most difficult. Therefore, we have used the Green procedure (Green, 1970), that is, an averaging of the correct item difficulties achieved by a student. Using the flexilevel algorithm and this scoring process, the overall reliability and validity of the adaptive testing procedure seems reasonably satisfactory as it yields coefficients that vary between .6 and .8 (i.e., alpha coefficients and parallel test coefficients).

In addition, we are making plans to contrast two additional adaptive routines so as to resolve what we perceive as a critical problem, namely, the critical zone performer. In any given training situation, there is a critical criterion zone, typically being between the 70th and 90th percent level which is stipulated as a requirement for the attaining of course mastery. If a student scores close or within this level (consider it being bounded by the standard error of measurement), then one should collect more information prior to judging this student as having achieved the objectives or in need of further remediation. At least two approaches can be considered to resolve this problem. The first is an obvious approach simply involving the presentation of an additional set of items for this zone; this is similar to a branching test. A more promising one, especially given the role of the computer, is Bock's (1972) procedure for item latent structure which makes use of the information contained in wrong alternative answers. The Bock model appears to us to be a far more preferable procedure in terms of ongoing large-flow training situations and it shall be evaluated during the coming year within the AF/AIS context.

Data relating to reduction in testing time indicates that only approximately 31 percent of the items are utilized if individualized entry and adaptive techniques are employed. This yields a 150 percent savings in testing time. The samples unfortunately, were extremely small and our group looks forward to a much more extensive validation study in the AIS military training situation. Similar savings are reported by Tam (1973) in his study of affective adaptive testing although modest ones were reported by Hedl (1971) in his intelligence testing. All in all, the results are sufficiently promising to extend the validation for these approaches as well as explore alternative designs within realistic training situations. These alternatives form the substance of the remainder of the paper.

Issues in Adaptive Testing

As an active reader and investigator in the adaptive testing area over the last eight years, one general observation comes to mind, namely, a classical psychometric approach emphasizing those cherished characteristics of excellence, improved reliability, validity, and consequential individual description, is limited in its systems and institutional view. In essence, our efforts have been to describe each and every individual in reliable, finegrain terms while recognizing the needs to improve the testing system. Given these broader insights, the purpose of this section will be to raise issues and possible alternatives as reflected by priorities concerning objectives for adaptive testing. There are three areas to be considered as reflected by these queries: (1) What are the possible purposes for adaptive testing? (2) What types of formal models might best be pursued for adaptive testing? and (3) How can our theoretical and procedural methods best be evaluated?

Purposes for Adaptive Testing

The tradition within psychometric research as well as test development has focused on descriptions and decisions concerning individuals. On the other hand, many institutions believe group differences in the testing process should be stressed since it is group data that form the basis of decision making. For example, in the current controversy concerning the contribution of schools and curriculum effects, Rakow (1974) argued that tests have been constructed to maximize on individual discriminations and to minimize group differences. Therefore it is not surprising that one finds no statistically significant group effects for schools or curriculums; the Coleman study (1968) or the Jencks follow-on study (1972) represent this type of outcome. Rakow argues that if one utilizes inter-class correlational techniques, one can find highly significant relationships of a subset of items which distinguish among groups. For adaptive tests that attempt to support large human organizations such as military training, this implies that classifying an individual concerning group membership and the characteristics of this group is of a high priority. This adaptive testing approach would utilize a branching item technique so as to lead to reliable alternative group classifications for an individual. Having achieved this, then the more conventional individual discrimination techniques could be applied. Obviously, the utilization of a flexilevel algorithm based on appropriate individual placement would be preferable. The point of such a two-stage model is to provide for more effective adaptation for group placement and ultimately for maximizing on institutional criteria rather than individual criteria alone. Simply, might it be better to find the correct group for an individual rather than know his "true score" on some ability dimension?

In turn, one can look at training systems and recognize that there is a trade-off between training load vs. standard error effects. In essence, as the training load absorbs more and more of the readily available resources, an improvement in the testing process with an associated reduction in standard error is superfluous since all the remaining individuals will have the same minimal treatment. In essence, each student is likely to spend long waiting times and not be able to pursue any kind of optimum course of instruction. Under such circumstances, it is therefore critically important to identify those individuals who can pursue self-study where appropriate. Moreover, it might also be highly important to have adaptive tests that better detect those individuals who seem to have aptitudes for transfer, so that when branched forward or back for review within a normal sequence of instruction, they will receive facilitating effects rather than negative ones.

In turn, as the training load on resources diminishes, one should expect the test length to increase so as to reduce errors of measurement. Thus, one can see that a systems approach to adaptive testing tends to reflect a far more dynamic procedure which might change the criteria, the test length, and the algorithms depending on the state of the training system.

Finally, to be optimally adaptive, one should recognize that our clientele and their institution basically do not understand the concepts, methods, or models of adaptive testing. To them, the quantification, especially as represented by our psychometric models, tends to defy understanding. Allow me to illustrate. MSU has been teaching a measurement course on base at NAS, Memphis. Two of the students were commanding officers of Navy technical training schools and have direct responsibility for supervising the measurement processes within these schools. After completing an eight-week course, each volunteered that they had, prior to the course, never understood any of the quantitative test item statistics or reports other than those concerning students passing or failing, the all important attrition rate. To be adaptive the system should provide the commanding officers, instructors, students, and other concerned people with verbal reports rather than quantitative reports; thus, a client-oriented product approach would vastly enhance the acceptance of adaptive testing. The work of Fowler (1969) with the MMPI successfully demonstrates that psychiatrists readily desire and understand verbal interpretations rather than quantitative reports of the 13 MMPI subscales. These observations about institutional effects hopefully will stimulate your interest in thinking about your clientele as well as your model when you formulate some of your priorities for future research. As cited in the introduction, adaptive testing research must be scholarly, diligent, and of the highest quality while reflecting a form of institutional adaptation which can be appreciated and supported by the clientele who provide the resource support for all research.

Psychometric Models for Adaptive Testing

Within the tradition of adaptive testing research, one reads numerous reports that focus on the comparative merits of alternative psychometric models for adaptive testing. It shall be the thesis of this section that pursuit of an optimal adaptive testing model is likely to be ineffective and the adaptive testing domain needs a strategy for identifying selection criteria that chooses among the many existing models. Optimization studies, especially from a formal point of view, have been pursued for the last 30 years in different contexts with surprisingly similar indifferent results. For example, during the 1940's many statisticians pursued within analysis of variance models the issue of optimal *a posteriori* mean difference tests. After better than a decade and a half of effort, John Tukey (1962) observed that one could not really argue for the one best *a posteriori* test because each varies according to the decision criteria of the investigator. In essence, it is the characteristic of the research which determines which one of the many tests is the most appropriate.

In turn, the area of mathematical learning models offers a similar finding. Within the context of research on the all-or-none vs. incremental learning processes during the early 1960's, one notes a flurry of research, all of which ended with the conclusion (Atkinson, et. al., 1965) that each mathematical learning model has a set of task characteristics which allows it to be optimal provided that the *a priori* task characteristics are sufficiently matched.

Recently a great deal of effort has gone into the investigation of adaptive instructional models from an optimization point of view. Generalized approaches include various regression models. While these regression models are clearly non-optimal, they have proven significantly successful in facilitating the process. On the other hand, fairly specific models, be these Markoff processes or dynamic programming structures, provide an elegant theoretical explanation (Hansen, et. al., 1973) but rarely fit the data or facilitate learning. Thus, one is led to the view that an array of models for the instructional area will be necessary in order to fit the rather diverse nature of the learning process.

Based on these examples, the proliferation of psychometric models for adaptive testing is likely to have limited productivity. Our efforts to focus on the criteria to be used for the selection of a given adaptive testing model and a better description of how to test the model's fit with the given behavioral phenomena would seem to be a more desirable direction in which to move.

Validation Procedures

As has been observed by each of the reviewers in this area, the amount of empirical work is modest at best. If one considers critical topics, namely, sample size and design techniques, one is even further impressed by our modest beginnings. For example, in reference to sample-size there are those such as Bock (personal communication) who would advocate that at least for his latent item structure model, a sample size of 2,000 students would be required. While pursuing some of the test data for the Air Force with a sample of 1,000 plus airmen, the groups were divided into samples of 200 each and then the usual reliability and validity analysis was performed. In addition, each sample was progressively aggregated into the next. It is fairly clear that the parameter convergence process. was still taking place after the sample size had increased to 800. Therefore, it can be argued that it is important to consider maximizing on sample size and to develop techniques by which both item and test parameters converge on their appropriate group and individual values.

In turn, our review of the designs for validation is consistent with that proposed by Tam (1973), namely, that one has to consider a within-test as well as a between-test validation procedure. This can be achieved simultaneously if one notes that one can present adaptive testing as a variation within total test procedure. In turn, this can be contrasted with a parallel form presentation. The two statistics, correlation between the two adaptive and total test scores and the correlations between the two parallel forms, yield a comprehensive representation of the validity. While this may seem excessive to some, such validation procedures provide more substantial empirical results which clearly indicate the justification for reducing total test items.

Summary

This review and reflection has run on in a rather extensive manner. Furthermore, it seems inappropriate to have reflections on reflections. Therefore, this summary will state a final point of view, namely, adaptive testing is sufficiently dynamic that multiple concepts and hypotheses can be incorporated in a design sequentially so as to determine their effect on the efficiency and effectiveness of the assessment process. This extensive review of a number of neglected topics should not be taken as a set of imperatives for research. Rather, these topics and suggestions can best be considered as potential variations within experimental designs of the future. They are offered to you under the assumption of collegial productivity and a firm commitment to the human and societal benefits from adaptive testing. Of all the evaluational techniques available to us at this time, adaptive testing offers that chance to humanize our assessment processes. Such an eventuality, especially in terms of shortening high-stress situations commonly found in testing, cannot be minimized in terms of its benefits.

BIBLIOGRAPHY

Atkinson, R. C., Bower, G. H., Crothers, E. J., *An Introduction to Mathematical Learning Theory,* John Wiley and Sons, Inc. 1965.

Bock, R. D. "Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories." *Psychometrika* 37 (March 1972): 29-51.

Coleman, J. S., et. al., *Equality of Educational Opportunity,* (Washington, D.C.: United States Government Printing Office, 1966).

Fowler, R. D. "The Current Status of a Computer Interpretation of Psychological Tests." *American Journal of Psychiatry* 125 (Supp. 1969): 21-27.

Green, B.F. "Comments on Tailored Testing." In W. Holzman, ed., *Computer-Assisted Instruction, Testing and Guidance.* New York: Harper and Row, 1970.

Hansen, D. N., Brown, B., Merrill, P., Tennyson, R., Thomas, D., Kribs, H. D. *The analysis and development of an adaptive instructional model(s) for individualized technical training – Phase 1,* Technical Report AFHRL-TR-72-50(1), Tallahassee, Fla.: CAI Center, Florida State University, 1973.

Hansen, D. N., Merrill P. F., Tennyson, R. D., Thomas, D. B., Kribs, H. D., Taylor, S. T. and James, T. G. *The Analysis and Development of an Adaptive Instructional Model(s) for Individualized Technical Training.* Technical Report for Contract No. F33615-71-C-1277, Air Force Systems Command. Tallahassee: Florida State University, 1973.

Hansen, D. N., Ross, S. *A Proposal to Study Additivity for Adaptive Instructional Treatments: A Theoretical Issue, Report for NPRDC, San Diego, Calif.* Memphis State University, 1975.

Hedl, J. J., Jr. *An Evaluation of a Computer-Based Intelligence Test.* Technical Report 21. Tallahassee: Florida State University CAI Center, 1971.

Jencks, C. S., "The Coleman Report and the Conventional Wisdom" in *On Equality of Educational Opportunity,* ed. by Frederick Mosteller and Daniel P. Moynihan, New York: Random House, Inc., 1972.

Rakow, E. A., "Evaluation of Educational Program Differences Via Achievement Test Item Difficulties", paper for American Educational Research Association, Chicago, Illinois, April, 1974.

Tam, P. T. "A Multivariate Experimental Study of Three Computerized Adaptive Testing Models for the Measurement of Attitudes toward Teaching Effectiveness." Unpublished Ph.D. dissertation, Florida State University.

Tukey, J. W. The Future of Data Analysis. *The Annals of Mathematical Statistics,* 33, 1-67. 1962.

COMPUTER ASSISTED TESTING: AN ORDERLY TRANSITION FROM THEORY TO PRACTICE

RICHARD H. MCKILLIP
VERN W. URRY
U.S. Civil Service Commission

The United States Civil Service Commission is responsible for examining applicants for Federal jobs throughout the world. It examines almost two million persons and makes about 200,000 placements annually.

The Commission's investment in computerized adaptive testing research and development is a significant one. This exciting and innovative program is currently budgeted at almost $200,000 per year. This expenditure comes at a time when Federal agencies' budgets are most austere and when resources are sorely needed to respond to the increasing challenges faced by conventional examining methods.

The Commission's investment in computerized adaptive testing is based primarily on the potential payoff in improved employee selection and placement. The large numbers of examinations and applicants makes computerized adaptive testing an economical, practical vehicle for improved measurement. The answer to attacks on tests in the employment situation is complex; the economic and social implications of this problem are enormous. Unquestionably, however, the greatest benefit both to the employer and to the employee lies in better measurement, not in less measurement. Every improvement in the selection and placement processes should contribute to the economic health of the employer, the psychological well being of the affected individual, and the welfare of society. Computer technology offers not only an opportunity to make significant improvements in employment decisions but also a better means of assessing the effects of such improvements.

While there are problems yet to be solved, computerized adaptive testing is well on the way to implementation.

As conventional approaches to test construction are modified in light of developments in latent trait theory, computerized adaptive testing becomes more and more feasible. The Rasch Model showed capabilities for computerized adaptive testing in the special case where all items discriminated equally and were unaffected by guessing. This special case was simply not practical to expect in available test items (Urry, 1970). Since item requirements for three-parameter logistic or normal ogive models can be met with existing items (Lord, 1970), computerized adaptive testing can be implemented. The implementation can be cost effective (i.e., the number of test items administered is substantially reduced vis-a-vis conventional testing) when certain rigorous item bank specifications can be met (Jensema, 1975). The determination that the item bank specifications can be met with existing items is contingent upon a new look at conventional item statistics and their relationship to model parameters. It has become apparent that the distortions caused by guessing result in severe underestimates, particularly of item discriminatory powers (Urry, 1975). Reliable estimates of parameters can now be made (Gugel, *et al*, 1975). An algorithm exists that will allow on-line computer-interactive item calibration (Schmidt & Urry, 1975).

Problems remain in tailoring test batteries to specific occupational requirements and in adequate coverage of job-related abilities. Of serious concern are the time and dollar respources that are needed for comprehensive measurement. The improved medium of presentation inherent in the hardware will facilitate resolution of these problems; for example, new item types and audio input possibilities.

Application of computerized adaptive testing in civil service examining has several desirable features.

Job relatedness. With multivariate test item banks, it is feasible to interpret scores on specific abilities in terms of differential occupational requirements. This then enables the employer to test a large number of abilities and to weight these abilities in accordance with their importance for success in specific jobs. The employer can array applicants across a large number of jobs and select in terms of priority, thus maximizing the utility of the selection process.

Standardized Examination Administration. Individual differences among administrators under conventional testing make error variance due to unstandardized administration largely unavoidable. Since administration procedures can be programmed under individualized testing, standard conditions can be better maintained.

Compromise of Examination Materials. Under computerized adaptive testing, examination questions are located in a central computer. No test booklets are used, therefore none can be taken from the examination room. As a result, the security of tests and test questions can be maintained more easily. Different individuals will receive different sequences of items, reducing the likelihood of cheating.

Improved Administrative Procedures. Test booklet printing, storage, and distribution costs become inconsequential.

Examination Scheduling. Tests can be administered on a walk-in basis since different tests can be administered simultaneously. The shortened testing time makes possible the administration of a multiple abilities battery in the time now required to examine for a single ability. Further, if selection is specific to a given position, individualized testing for the required abilities can be accomplished in a manner that minimizes the time of testing while maximizing the job relatedness of a final weighted score.

Power Conditions of Examination. Tests of ability should be power tests. However, due to administrative considerations, i.e., scheduling, space restrictions, etc., conventional tests of ability are usually speeded to a certain degree. Under computerized adaptive testing, the power conditions required by this type of test can be ensured.

Test-Taking Motivation. Test-taking motivation and, consequently, test performance may be impaired when the level of difficulty of the examination material is inappropriate to the level of ability of the examinee. In conventional testing, the examination is constructed for an entire population. This method of construction necessarily leads to inappropriate question difficulties when a conventional test is presented to a given examinee. In computerized adaptive testing, the difficulty level of the questions is matched to the level of ability of the examinee.

Improving Examinations. The current conventional testing technology is the product of more than fifty years of research and development. Substantial improvements have been less frequent with the passage of time. This calls for a rather dramatic change in testing procedure. At present, the appropriate change would be towards an individualized testing technology. Certainly greater experimental control and a thorough monitoring of the measurement process is made possible through the aid of this new medium.

Improving Personnel Decisions. When a computer interactive network has been established for individualized testing, one has necessarily established a vast data accession network to effect immediate evaluation of the personnel decision making process. Optimization in the decision-making process is the natural extension of events when many sources of information are available to a central computer and are readily accessible for analysis by the personnel researcher and personnel specialist.

It appears, at this time, that computerized adaptive testing research has progressed to the point where implementation will be feasible. In Fiscal Year 1976, a comprehensive cost analysis will be undertaken. Preliminary estimates are favorable. For example, computer connect time in testing in one ability area now costs less than forty cents per examinee. It is reasonable to expect that cost to drop as the program progresses. Current plans call for fully operational computerized adaptive testing by 1980. At that time, it is expected that the examination for most entry-level professional and administrative jobs will include a test battery administered in the computerized adaptive system. Approximately 200,000 applicants currently file for these jobs. It will take until 1980 to get ready for an examination of this scope and number of participants.

My colleagues this morning will address some of the progress we have made in solving technical problems associated with the program.

REFERENCES

Gugel, J. F., Schmidt, F. L. & Urry, V. W. *Effectiveness of the ancillary estimation procedure.* Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

Jensema, C. J. *Bayesian tailored testing and the influence on item bank characteristics.* Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

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.

Schmidt, F. L. & Urry, V. W. *Item parameterization procedures for the future.* Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

Urry, V. W. A Monte Carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). *Dissertation Abstracts Internatinal,* 1971, 31, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, 34, 253-269.

Urry, V. W. *The effects of guessing on parameters of item discriminatory power.* TN-75-2 Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, 1975.

A FIVE-YEAR QUEST: IS COMPUTERIZED ADAPTIVE TESTING FEASIBLE?

VERN W. URRY
U.S. Civil Service Commission

Five years of research on the feasibility of computer assisted testing has attempted to answer four extremely significant questions: (1) What types of items are required for effective computerized adaptive testing? (2) Do these types of items exist in sufficient number to measure important abilities adequately? (3) Can estimates of the item parameters be obtained that are sufficiently reliable to be used successfully in a computerized adaptive testing algorithm? and (4) Is there an efficient and accurate adaptive algorithm for computerized testing?

In answer to the first question, "What types of items are required for effective computerized adaptive testing?", the development of specifications for effective item banks or item pools for computerized adaptive testing was begun about five years ago (Urry, 1970). These specifications were written with reference to the three parameters of the normal ogive model (Lord & Novick, 1968) and the logistic model (Birnbaum, 1968). At that time, they included requirements for a minimum of 100 items with item discriminatory powers (the a_i) of at least .80, with item difficulties (the b_i) evenly distributed on the interval from −2.00 to 2.00, and with item coefficients of guessing (the c_i) of .25 as a maximum. Some research was later completed (Jensema, 1974; Urry, 1974*b*) indicating that the maximum value for the c_i could be set as high as .30 with item bank effectiveness still maintained.

In these studies, an item bank was adjudged effective when computerized adaptive testing required fewer items than conventional paper and pencil testing to attain the same level of reliability. The specifications were arrived at through model sampling and simulation techniques. The concern was the capability of the 3-parameter models for the specific purpose of computerized adaptive testing. After model capabilities were adequately explored, there remained the empirical question, "Do these types of items exist in sufficient number to measure important abilities adequately?"

At first glance, it might have appeared that the requirement for item discriminatory powers of .8 or greater was unreasonably high given the usual test item because an item discriminatory power of .8 corresponds to a biserial correlation of .62 between the item and latent ability. In the experience of most psychometricians this would seem an impossible specification to meet, because the usual item-test biserial correlations tend to be much lower than this specified value. However, the impossibility exists only if the attenuating effects of guessing on conventional indicants of item discriminatory power are not fully understood. These effects mask the true discriminatory power of multiple-choice items to a marked degree, and they are still largely unappreciated.

In order to illustrate these effects, equations were derived for the point-biserial (Urry, 1974*a*) and the biserial (Urry, 1975) correlations between multiple-choice items and latent ability. The equation for the point-biserial correlation was derived as

$$\rho_{i'\theta} = \frac{(1 - c_i)\,\rho_{I\theta}\,\phi(\gamma_i)}{\sqrt{P_i'\,Q_i'}}$$

(Urry, 1974*a*, eq. 15); (1)

and the derivation of the biserial correlation resulted in

$$\rho_{I'\theta} = \frac{(1 - c_i)\,\rho_{I\theta}\,\phi(\gamma_i)}{\phi(\gamma_i')}$$

(Urry, 1975, eq. 6). (2)

In these equations, a prime was used to indicate that the given term was affected by guessing. Definitions were as follows:

c_i the item coefficient of guessing, is the lower asymptote of the regression of the binary item on latent ability;

$\rho_{I\theta}$ is the biserial correlation, unaffected by guessing, between the binary item and latent ability;

γ_i is the baseline value of the item distribution *N*(0,1) above which the probability of (or proportion) knowing the correct response occurs;

$\phi(\gamma_i)$ is the height of the ordinate at γ_i;

P_i' is the probability of (or proportion) passing a multiple-choice item;

Q_i' or $1 - P_i'$, is the probability of (or proportion) missing a multiple-choice item;

γ_i' is the baseline value on the distribution $N(0,1)$ above which the probability of (or proportion) passing, viz. P_i', occurs:

$\phi(\gamma_i')$ is the height of the ordinate at γ_i'.

The difference between the probability of (or proportion) knowing the correct response to an item, viz.,

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{\gamma_i}^{\infty} exp \left[\frac{-t^2}{2} \right] dt, \tag{3}$$

and the probability of (or proportion) passing a multiple-choice item, viz.,

$$P_i' = c_i + (1 - c_i) P_i, \tag{4}$$

is to be duly noted. As a consequence, it is known that γ_i is equal to γ_i' only when c_i is zero. When guessing is effective (or, synonomously, c_i is not zero), neither γ_i and γ_i' nor $\phi(\gamma_i)$ and $\phi(\gamma'_i)$ are equal. Further, when guessing is effective, γ_i', as a baseline value, is unlike γ_i which divides the item distribution meaningfully on the basis of success on the item. Notice that for c_i equal to zero, equation (2) indicates the equality of $\rho_{I\theta}'$ and $\rho_{I\theta}$. Otherwise the distinction between these two coefficients is to be kept clearly in mind. Since item discriminatory power is defined by the normal ogive model as

$$a_i \equiv \frac{\rho_{I\theta}}{\sqrt{1 - \rho_{I\theta}^2}} \;, \tag{5}$$

it is totally inappropriate to substitute estimates of $\rho_{I\theta}'$ for $\rho_{I\theta}$ in equation (5) to estimate a_i. When guessing is effective or when the items are of a multiple-choice variety, this procedural error adversely affects computerized adaptive testing.

The derived equations for the point-biserial and biserial correlations were used to illustrate the attenuating effects of guessing on these conventional indicants of item discriminatory power. In the procedure, the item coefficient of guessing is usually set at some meaningful value, say, the reciprocal of the number of alternatives for a multiple-choice question; and for this fixed value of c_i, the equations are evaluated to map the levels of a_i and b_i onto the planes defined by the coordinates, the point-biserial correlation and the p-value, or the biserial correlation and the p-value. In Figure 1, the levels of a, viz., .8, 1.0, 1.2, 1.4, 1.6, 2.0, and 3.0, and the levels of b, viz., 2.0, 1.6, . . . , -2.00, have been mapped onto the plane defined by the population point-biserial correlation and the population proportion passing or p-value for c equal to .20. When c is fixed at .20, the effectiveness of guessing is roughly equivalent to the level typical of 5-alternative items. Since the biserial correlation (unaffected by guessing) between the item and latent ability is defined as

$$\rho_{I\theta} \equiv \frac{a_i}{\sqrt{1 + a_i^2}} \tag{6}$$

in the normal ogive model, the levels of a portrayed in Figure 1, viz., .8, 1.0, 1.2, 1.4, 1.6, 2.0 and 3.0, correspond to item ability biserials of .62, .71, .77, .81, .85, .89, and .95. Notice then the apparent paradox. For example, an item which has an item-test point–biserial correlation of .11 with a p-value of .22 is indicated to have an item discriminatory power, a_i, of 3.00 or a $\rho_{I\theta}$ of .95. The astonishing paradox is due to the attenuating effect of guessing. In Figure 2, identical levels of a and b have been mapped onto the plane defined by the population biserial correlation and the population proportion passing or p-value, again, for c fixed at .20. While the attenuating effect is less pronounced for the biserial correlation relative to the point-biserial correlation, it is most severe for difficult items. For example, a five-alternative multiple-choice item with an item-test biserial correlation of .17 and a p-value of .22 is indicative of an item discriminatory power of 3.0 or an item-ability biserial of .95 and an item difficulty of 2.00. What would happen if the procedural

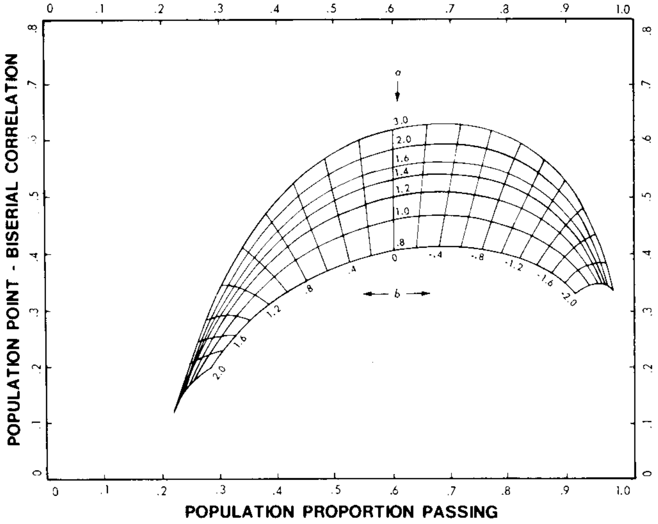


Figure 1. Relationship between conventional and normal ogive item parameters when the coefficient of guessing (c) equals .20.

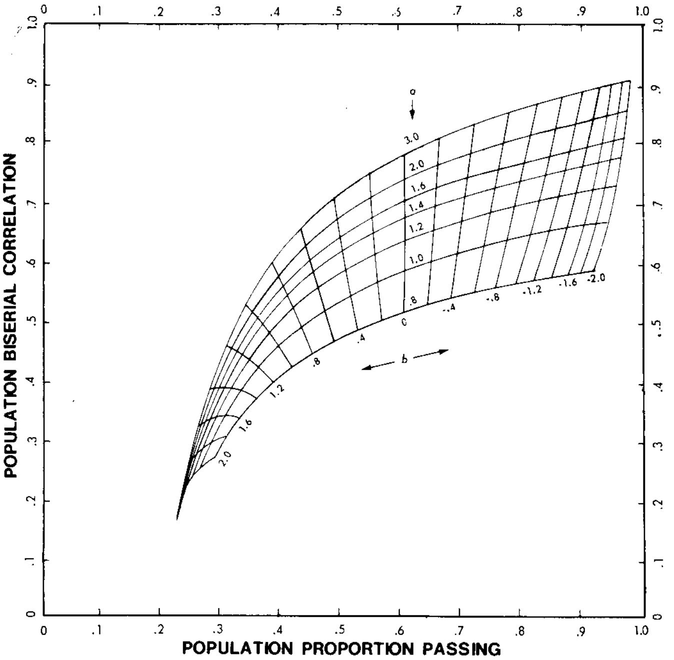


Figure 2. Relationship between conventional and normal ogive item parameters when the coefficient of guessing (c) equals .20.

error alluded to earlier were committed in connection with this interesting case? It will be recalled that the error involved the misuse of $\rho'_{I\theta}$ in equation (5). In this instance, a_i would have been erroneously estimated as .17 when the true value was 3.00. Obviously, gross errors of this nature render computerized adaptive testing less efficient than it should normally be. If the data point defined by the item-test point-biserial or biserial correlation and the p-value is plotted on one of these maps or charts, the corresponding values of a_i and b_i for the given item can be interpolated from the grid system that identifies the various levels of a_i and b_i. For reliable total tests[1] and large samples, the interpolated values of a_i and b_i approximate the true parameters and allow the researcher (1) to identify items appropriate for the purpose of computerized adaptive testing and (2) to assess the efficacy of a given set of appropriate items for the purpose of computerized adaptive testing by comparing the obtained interpolated values with the specifications for item bank effectiveness. When the specifications are met, improved reliability per item used is assured for computerized adaptive tests relative to conventional tests. However, the number of items required in computerized adaptive testing relative to conventional testing can be markedly reduced when the a_i appreciably exceed the minimum value of .80, the b_i are widely and evenly distributed, and the c_i are maintained at low values.

Experience has shown (Jensema, 1972; Urry, 1974*b*) that roughly one-third of the items in the usual aptitude or ability test survive this screening for appropriateness. Moreover, item discriminatory powers have been frequently found to exceed 2.0 in value.

After it was ascertained that sets of items could be found that would satisfy the specifications for effective item banks, there remained the important question, "Can estimates of the item parameters be obtained that are sufficiently reliable to be used successfully in a computerized adaptive testing algorithm?" In answer to this question, a relatively rapid and inexpensive item-analytic procedure was developed (Urry, in press–*a*). It has been programmed and is currently available for use on several computers. The output of the program is an item analysis yielding ancillary estimates for a_i, item discriminatory power; b_i, item difficulty; and c_i, item coefficient of guessing.

Estimates of the parameters a_i, b_i, and c_i are obtained by an iterative minimum χ-square procedure. The procedure consists of two stages that differ only with respect to the particular measure used for manifest ability. In the first stage, the distribution of manifest ability is represented by corrected raw scores where the item being parameterized is omitted from the scoring. In the second stage, the distribution of manifest ability is represented by Bayesian modal estimates of ability (Samejima, 1969). Generally, Bayesian modal estimates of ability more closely approximate the distribution of latent ability than does the distribution of corrected raw scores. Therefore, the second stage constitutes a refinement on the first stage. In both stages the procedure iterates item by item through values of c_i to obtain pairs of a_i and b_i consistent with large sample estimates of the item-manifest ability point-biserial correlation and the item p-value. This allows the generation of various item characteristic curves (ICC's). The ICC's are then compared with the regression of the binary item on manifest ability. The ICC that best fits this regression, as indicated by the minimum χ-square, is given by the set of approximations – $\hat{\hat{a}}_i$, $\hat{\hat{b}}_i$, and $\hat{\hat{c}}_i$. The approximations are then corrected for characteristics of the particular sample of items being parameterized to obtain "ancillary estimates" – $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$. Ancillary estimation as a generic method was developed by Fisher (1950). The ancillary corrections improve the efficiency of the estimates.

The procedure has been evaluated through model sampling and simulation techniques. In particular, two parameterization samples, one of 2,000 and one of 3,000 cases, were generated from the logistic model using specified, and hence known, item parameters. The data were then analyzed by the procedure, and the resulting estimates were compared to the known parameters for each

[1] As total test reliability decreases, the approximations for the parameters a_i systematically underestimate the true values of a_i.

of the samples. Specifically, root mean square errors (RMSE's), i.e.

$$\left\{ \sum_{i=1}^{m} (\hat{a}_i - a_i)^2 \, m^{-1} \right\}^{1/2}, \qquad \left\{ \sum_{i=1}^{m} (\hat{b}_i - b_i)^2 \, m^{-1} \right\}^{1/2}, \text{ and}$$

$$\left\{ \sum_{i=1}^{m} (\hat{c}_i - c_i)^2 \, m^{-1} \right\}^{1/2}$$

, were obtained. These measures of deviation are given in Table 1 for the two parameterization samples and stages. Notice that the particular RMSE indicated by a given equation tends to decrease with stages. This is an indication of improved efficiency due to ancillary corrections. For the final stage ancillary estimates, these deviation measures were .242, .123 and .056 for the 2000 case sample, and .228, .148, and .056 for the 3000 case sample. For 100-item parameterization tests, these data indicated that 2,000 cases were sufficient for the effective use of the procedure. Correlations were also computed between the estimates and the known parameters, i.e., $r_{\hat{a}a}$, $r_{\hat{b}b}$, and $r_{\hat{c}c}$. These correlations are provided in Table 2 for the two parameterization samples and stages. Notice that there is a tendency for each correlation to increase with stages as predicted given that the ancillary corrections improve efficiency of estimation. For the final stage ancillary estimates, the correlations were .915, .996, and .764 for the 2,000 case sample, and .918, .997, and .760 for the 3,000 case sample. Since the ranges of the a_i and c_i were somewhat restricted, these correlations are very respectable. The results of these comparisons between the estimates and the known parameters indicated the merit of the item-analytic procedure.

The ancillary estimation procedure was further evaluated using simulation techniques. In particular, testing was conducted using a Bayesian algorithm developed by Owen (1969). Samples of 100 cases each were generated for computerized adaptive testing using 100 items with known item parameters. In the generation process, values of θ, the ability parameter, are sampled randomly from $N(0,1)$ and are also known. As a result, estimates of the ability obtained under computerized adaptive testing could be correlated with known ability. Comparisons of correlations, $r_{\hat{\theta}\theta}$, were made across three conditions of computerized adaptive testing where (1) the known item parameters, (2) the ancillary estimates of the item parameters based on the 2,000 case sample, and (3) the ancillary estimates of item parameters based on the 3,000 case sample were used in the algorithm. The appropriateness of the use of the ancillary estimates could be evaluated, therefore, by comparing the results obtained for the last two conditions with those

TABLE 1

Root Mean Square Errors for Estimates by Parameterization Samples and Stages

Sample Size	Parameterization Stage	Root Mean Square Error		
		$\left(\sum_{i=1}^{m}\left\{\hat{a}_i - a_i\right\}^2 m^{-1}\right)^{1/2}$	$\left(\sum_{i=1}^{m}\left\{\hat{b}_i - b_i\right\}^2 m^{-1}\right)^{1/2}$	$\left(\sum_{i=1}^{m}\left\{\hat{c}_i - c_i\right\}^2 m^{-1}\right)^{1/2}$
2000	Corrected Raw Score:			
	Approximation	.309	.181	.077
	Ancillary Estimate	.283	.120	.067
	Bayesian Modal:			
	Approximation	.269	.150	.061
	Ancillary Estimate	.242	.123	.056
3000	Corrected Raw Score:			
	Approximation	.308	.139	.081
	Ancillary Estimate	.253	.135	.073
	Bayesian Modal:			
	Approximation	.252	.109	.059
	Ancillary Estimate	.228	.148	.056

TABLE 2

Correlations Between Estimates and Known Parameters by Parameterization Samples and Stages

Sample Size	Parameterization Stage	Correlation		
		$r_{\hat{a}a}$	$r_{\hat{b}b}$	$r_{\hat{c}c}$
2000	Corrected Raw Score:			
	Approximation	.876	.996	.651
	Ancillary Estimate	.873	.996	.668
	Bayesian Modal:			
	Approximation	.909	.996	.754
	Ancillary Estimate	.915	.996	.764
3000	Corrected Raw Score:			
	Approximation	.884	.996	.611
	Ancillary Estimate	.895	.996	.616
	Bayesian Modal:			
	Approximation	.914	.997	.752
	Ancillary Estimate	.918	.997	.760

obtained for the first. In Table 3, the results are summarized for each of the conditions of testing.

Further explanation, however, is in order before proceeding to an interpretation of these results. When compared with conventional testing procedures, computerized adaptive testing can lead to a substantial reduction in the number of items required to obtain a given degree of validity. Therefore, the concern was not only with the validity obtained but also with the economy in items observed in obtaining the given validity. Control over the validity of computerized adaptive testing is direct. When an individual is being evaluated, the standard error of the estimate of ability is available at any stage in the sequence. Validity, over individuals, is controlled by terminating the

TABLE 3

Validity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for Tailored Testing to Various Termination Rules Where the Item Parameters Were Known or Estimated

Termination Rules						Item Parameters Estimated in a Sample of:			
#	σ_ϵ	$\rho^2_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	Parameters Known		2,000 Cases		3,000 Cases	
				$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
1	.5477	.70	.84	.84	2.7	.83	2.0	.84	2.3
2	.5000	.75	.87	.85	3.2	.86	2.7	.86	2.6
3	.4472	.80	.89	.89	3.9	.89	3.4	.88	3.2
4	.3873	.85	.92	.91	4.7	.90	4.0	.90	4.0
5	.3162	.90	.95	.94	6.6	.92	5.4	.93	5.6
6	.2828	.92	.96	.96	8.2	.94	6.7	.93	7.1
7	.2449	.94	.97	.96	10.8	.95	9.1	.94	9.6
8	.2236	.95	.97	.96	13.3	.95	11.1	.95	11.9

individual sequences at a common value for the standard error of the estimate of ability. In the study, eight such termination rules were designated. These rules are identified in columns 1 and 2 of Table 3 and specify that the standard error of the estimate of ability, σ_ϵ, was equal to or less than (1) .5477, (2) .5000, (3) .4472 (4) .3873, (5) .3162, (6) .2828, (7) .2449 and (8) .2236, respectively, over all individuals. Given σ_ϵ for any termination rule, synonomous rules may be generated through

$$\rho^2_{\hat{\theta}\theta} = 1 - \sigma^2_\epsilon \tag{7}$$

and

$$\rho_{\hat{\theta}\theta} = \sqrt{1 - \sigma^2_\epsilon} \tag{8}$$

for the expected reliability and validity, respectively. These synonomous rules are given in column 3 and 4. The validities of column 4 may then be compared with obtained validities. Eight estimates of ability satisfying these rules were obtained for all cases. Obtained validities were indexed by the correlations between known ability and estimated ability $r_{\hat{\theta}\theta}$, for specified termination rules as appropriate to the testing condition. As the termination rule becomes more stringent, the obtained validities given in columns 5, 7, and 9 increase and compare very closely with expected validities given in column 4. Additionally, the average numbers of items required, the $\bar{n}$, given in columns 6, 8, and 10 also increase as the termination rule becomes more stringent. Notice that the $\bar{n}$ at each termination rule differ only slightly across testing conditions. Since the results were almost identical across testing conditions, the item-analytic procedure appeared very appropriate in computerized adaptive testing applications. Consequently, ancillary estimates of the item parameters based on more than 2,000 cases and 100 items were strongly recommended for use in computerized adaptive testing.

Further research in evaluating the item-analytic procedure has been accomplished for varying numbers of cases and items (Gugel et. al., 1975), and more detailed recommedations regarding the use of the procedure will be given later in the conference.

As it turned out, the last significant question, "Is there an efficient and accurate adaptive algorithm for computerized testing?" could have been answered in the affirmative as early as 1969. The important event was the publication of an Educational Testing Service research bulletin, "A Bayesian Approach to Tailored Testing", by Roger J. Owen. Subsequent research (Urry, 1971, 1974*b*, in press-*a*; Jensema, 1972, 1974, 1975) has shown the efficiency and accuracy of the algorithm. For example, it is possible to construct some 2,000 computerized adaptive tests in some 17 minutes of central processor unit time, and the precision of measurement can be accurately controlled with termination rules.

In summary, we now find that: (1) the specifications for effective item banks have been developed, (2) these specifications can be met for a number of significant abilities, (3) efficient procedures exist for the reliable estimation of parameters, and (4) an efficient computerized adaptive testing algorithm is available to conduct the actual testing. All the necessary prerequisites for the success of computerized adaptive testing are therefore now in evidence. At this juncture, the feasibility of computerized adaptive testing can be realistically assessed, and this realistic assessment is decidedly and resoundingly affirmative in nature. At present, computerized adaptive testing appears to have a future without parallel in the literature of psychological measurement.

REFERENCES

Birnbaum, A. Part 5. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Fisher, R.A. Contributions to mathematical statistics. New York: John Wiley & Sons, 1950.

Gugel, J. F., Schmidt, F. L. & Urry, V. W. Effectiveness of the ancillary estimation procedure. Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college test battery. Unpublished doctoral dissertation, University of Washington, 1972.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, 34, 757-766.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Paper presented at the conference on Computerized Adaptive Testing, Washington, D.C.: June 1975.

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

Owen, R. J. A Bayesian approach to tailored testing. Research Bulletin 69-92. Princeton, N.J.: Educational Testing Service, 1969.

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

Urry, V. W. A Monte Carlo investigation of logistic mental test models. (Doctoral dissertation, Purdue University, 1970). *Dissertation Abstracts International,* 1971, 31, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Individualized testing by Bayesian estimation. Seattle: Bureau of Testing, University of Washington, 1971. (Duplicated Report)

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, 34, 253-269. (a)

Urry, V. W. Computer-assisted testing: calibration and evaluation of the verbal ability bank (TS-74-3). Washington D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, December 1974. (b)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In press-a.

Urry, V. W. The effects of guessing on parameters of item discriminatory power. (TN-75-2) Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, May 1975.

EFFECTIVENESS OF THE ANCILLARY ESTIMATION PROCEDURE [1]

JOHN F. GUGEL, FRANK L. SCHMIDT, AND VERN W. URRY
U.S. Civil Service Commission

Urry (1974a) has presented a graphic method to provide approximations for the item parameters of the normal ogive and Birnbaum logistic three-parameter latent trait models. This method has since been further developed (Urry, 1975) to provide a more accurate computational procedure for estimating the three parameters, a_i (item discriminatory power), b_i (item difficulty), and c_i (item coefficient of guessing). Programmed for the computer, this technique produces parameter estimates quickly and inexpensively.

Initial studies of this procedure employed large sample sizes (N=2000 and 3000 cases) and a relatively large number of items (n=100). Under these conditions, the procedure produces very accurate parameter estimates (Urry, 1975). We are now in a position to examine the effects of reduced numbers of cases and items on error in the parameter estimates and on the accuracy of tailored testing using those estimates. It is known *a priori*, of course, that reduction in either the number of cases or the number of items will, other things being constant, tend to increase estimation errors. But it is not known at present how large or practically significant such increases would be. The present study, exploratory in nature, is addressed to these questions.

METHOD

Based on suggestions by Lord (1968, p. 1016) and the results of the previous study by Urry (1975), it was decided to allow the number of items to vary from 50 to 100 and the number of cases to range from 500 to 2000. The initial 100-item bank, from which the smaller banks were later selected, was characterized by a_i values ranging uniformly from .80 to 2.20, b_i values distributed uniformly from −1.9 to +1.9, and c_i values from .02 to .24, also uniform in distribution. These parameter values are not different from what one might reasonably expect to find empirically given prescreening of items (Urry, 1974*a*; Jensema, 1972). In the reduced item samples, the a_i values were chosen in equal steps from .80 to 2.20. For example, there were five levels of a_i for the 50-item test and ten for the 100-item test. Ten values of b_i in equal steps between −1.9 and 1.9, inclusive, were arranged within each level of a_i. (an exception was the 55-item test, which had eleven values of b_i in equal steps between −1.9 and 1.9, inclusive, within each of its a_i values.) For different levels of a_i, items were matched on b_i values. The c_i values ranged from .02 to .24 in equal steps, irrespective of a_i and b_i. Values of θ, representing simulated subjects, were sampled randomly from $N(0,1)$. Then for each θ, the simulation procedure described by Urry (1975 was used to generate a vector of responses (1 = correct; 0 = incorrect) for the item bank in question using the known item parameters. Parameter estimation was then carried out using this simulated data.

Two indices were used to evaluate the parameter estimates relative to the known parameters. First, the root mean square error (RMSE) was computed for the estimated parameters. The formula for this statistic, is;

$$\text{RMSE} = \sum_{1}^{n}\left(\frac{(\rho - \hat{\rho})^2}{n}\right)^{1/2} \tag{1}$$

where the p = known values of a_i, b_i, c_i, or $\rho_{I\theta}$, and
n = number of items involved in the particular analyses.

Second, Pearson correlations between the known and estimated parameters were computed, i.e., $r_{p\hat{p}}$.

To illustrate the effects of error in the parameter estimates on the accuracy of tailored testing, Owen's (1968) algorithm was employed. Specifically, tailored testing was carried out on 100 simulated subjects using first the known item parameters and then item parameter estimates obtained on 1000 cases and 60 items. To increase the number of items used in tailored testing to a more realistic level, another identical set of 60 items was parameterized on a separate, independent group of 1000 simulated subjects, and these "items" were combined with the original 60 to produce a bank with 120 items. In the case of the known parameters, both 60-item sets were entered into the tailored testing bank. The known parameters in this bank were used to generate the response vectors of the 100 simulated subjects, and these vectors in turn, were used in the tailored testing. Correlations between estimated and actual θ were computed at each of eight termination rules for each condition of testing. This allowed a comparison of correlations across the conditions of testing, i.e., where (1) known or (2) estimated item parameters were used in the tailoring process.

[1] Computer processing for this study was done at the University of Maryland Computer Science Center in conjunction with graduate work by John Gugel. Arrangements for computer time were made by Professor Charles Johnson of the Department of Measurement and Statistics, College of Education, University of Maryland.

RESULTS AND DISCUSSION

Results produced by the parameterization procedure for varying combinations of sample size and number of items are shown in Tables 1 and 2. In both tables, "Raw Score Estimates" refer to the parameter estimates prior to application of the ancillary correction procedure, and the columns headed "Final Estimates" refer to estimates after application of the corrections. Table 1 includes the S.E. for $\rho_{I\theta}$, the correlation between the continuum underlying the item and θ, as well as for a_i, b_i, and c_i. "Lost items" are those for which the estimation procedure did not converge because of insufficient cases in the tails of the distribution.

Looking at the S.E.'s for the final estimates in Table 1, it can be seen that, in general, decreasing both sample size and number of items results in increased RMSE's. This effect appears to be more pronounced for a_i than for the other parameters. Moving from 50 to 60 items (sample size constant) appears to produce marked reductions in error for a_i, but beyond this, improvements in accuracy with increases in number of items are smaller. The b_i and c_i were estimated rather accurately throughout the range of both independent variables, although variation in sample size and number of items did have the expected effect. The last column in Table 1 reveals a tendency for items to begin to fail to converge during parameter estimation when sample size is dropped as low as 500. Sample size appears more crucial in this respect than number of items. Correlations between final parameter estimates and actual parameters, shown in Table 2, also pattern themselves as expected, within the limits of sampling error. In examining these correlations, one should bear in mind that in the case of $\hat{a}_i$ and to a lesser extent $\hat{c}_i$, restriction in range is operating to lower the tabled values. The items parameterized contained no values of a_i lower than .80. This value of a_i corresponds to a biserial correlation of .62 between the item and latent ability. Past studies (Jensema, 1972; Urry, 1974*b*) have shown that only about one third of the items in conventional tests have a_i values this large. No c_i greater than .24 were included; in practice c_i does exceed .24, although the range restriction here is probably less severe than in the case of a_i.

Results of simulated tailored testing using known parameters and parameters estimated on a sample of 1000 with 60 items are shown in Table 3. The eight termination rules, expressed as the standard error of estimate ($\sigma_{\hat{\epsilon}}$) are seen in column 2. Column 3 translates these values to reliability coefficients for $\hat{\theta}$, based on the relationship

$$\rho^2_{\hat{\theta}\theta} = 1 - \sigma_{\hat{\epsilon}}^{\,2} \tag{2}$$

TABLE 1

Root Mean Square Errors (RMSE)
Before and After all Corrections

		Raw Score Estimates				Final Estimates				
		RMSE				RMSE				Lost
Items	Cases	a_i	b_i	c_i	$\rho_{I\theta}$	a_i	b_i	c_i	$\rho_{I\theta}$	Items
50	2000	.283	.124	.086	.043	.395	.137	.064	.053	0
50	1000	.292	.193	.097	.053	.326	.209	.078	.059	1
50	500	.370	.164	.097	.067	.472	.259	.077	.064	0
55	2000	.385	.195	.091	.061	.308	.150	.057	.053	0
55	1000	.352	.194	.101	.050	.315	.124	.071	.050	0
55	500	.281	.185	.098	.054	.403	.227	.086	.065	4
60	2000	.321	.204	.091	.056	.253	.140	.065	.040	0
60	1000	.343	.231	.089	.059	.322	.144	.062	.044	0
60	500	.360	.194	.080	.070	.342	.179	.068	.062	0
70	2000	.272	.131	.095	.041	.225	.166	.067	.040	1
70	1000	.324	.189	.095	.054	.273	.174	.074	.045	1
70	500	.386	.197	.096	.072	.351	.187	.083	.058	4
80	2000	.266	.141	.092	.046	.214	.150	.072	.039	1
80	1000	.259	.178	.092	.048	.261	.166	.073	.047	1
80	500	.319	.224	.091	.063	.311	.229	.079	.048	6
90	2000	.297	.180	.094	.049	.244	.149	.069	.036	0
90	1000	.341	.171	.089	.051	.304	.140	.072	.044	0
90	500	.316	.184	.094	.056	.283	.144	.086	.049	2
100	2000	.290	.138	.085	.049	.223	.131	.056	.036	0
100	1000	.286	.137	.088	.052	.240	.162	.062	.039	0
100	500	.354	.189	.100	.061	.276	.161	.083	.047	5

TABLE 2

Correlations–Known Parameters vs. Estimated Parameters
Before and After All Corrections

Items	Cases	Raw Score Estimates			Final Estimates		
		$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$	$r_{a\hat{a}}$	$r_{b\hat{b}}$	$r_{c\hat{c}}$
50	2000	.846	.999	.599	.849	.997	.636
50	1000	.888	.992	.429	.908	.990	.492
50	500	.745	.993	.428	.780	.989	.454
55	2000	.731	.995	.488	.891	.995	.646
55	1000	.758	.995	.428	.870	.995	.546
55	500	.850	.992	.387	.824	.990	.376
60	2000	.828	.996	.491	.899	.997	.630
60	1000	.771	.994	.546	.842	.995	.588
60	500	.768	.994	.626	.801	.995	.668
70	2000	.834	.997	.471	.922	.997	.632
70	1000	.813	.996	.468	.828	.996	.521
70	500	.715	.993	.464	.813	.995	.449
80	2000	.873	.996	.535	.914	.997	.574
80	1000	.850	.994	.465	.879	.993	.550
80	500	.839	.991	.410	.823	.989	.502
90	2000	.861	.996	.483	.871	.996	.568
90	1000	.757	.995	.518	.847	.995	.547
90	500	.804	.995	.447	.874	.993	.418
100	2000	.837	.997	.539	.877	.998	.690
100	1000	.843	.996	.470	.863	.996	.627
100	500	.741	.993	.344	.824	.994	.420

The square root of this value is $\rho_{\hat{\theta}\theta}$, the correlation between the latent ability estimates ($\hat{\theta}$) and actual latent ability (θ). Validity coefficients of this sort are given in columns 4, 5, and 7. Those in column 4 are theoretical validities based solely on the termination rule chosen. Those in column 5 were obtained by correlating the $\hat{\theta}$ produced using the known item parameters with known θ. As expected they are essentially identical to the predicted theoretical validities. Those in column 7 were obtained by correlating the $\hat{\theta}$ produced using the parameter estimates with the known θ. As expected, they are somewhat lower than those in columns 4 and 5, but it can be noted that, as

TABLE 3

Validity Coefficients ($r_{\hat{\theta}\theta}$), and Average Number of Items ($\bar{n}$) Required for
Tailored Testing to Various Termination Rules Where the Item
Parameters Were Known or Estimated

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Termination Rules			Parameters Known		Parameters Estimated	
#	σ_ϵ	$\rho^2_{\hat{\theta}\theta}$	$\rho_{\hat{\theta}\theta}$	$r_{\hat{\theta}\theta}$	$\bar{n}$	$r_{\hat{\theta}\theta}$	$\bar{n}$
1	.5477	.70	.84	.864	2.43	.792	2.26
2	.5000	.75	.87	.904	3.31	.821	2.89
3	.4472	.80	.89	.932	4.00	.821	2.89
4	.3873	.85	.92	.935	4.91	.864	3.70
5	.3162	.90	.95	.955	7.03	.895	5.30
6	.2828	.92	.96	.962	8.77	.921	6.57
7	.2449	.94	.97	.969	11.77	.942	8.91
8	.2236	.95	.97	.975	14.51	.952	11.12

the termination rule becomes more stringent, the discrepancy decreases. At the most stringent termination rule, the validity of the $\hat{\theta}$ derived using the parameter estimates is only .023 lower than that based on the known parameters. The reliabilities of the two $\hat{\theta}$'s at this termination rule are .95 and .91, respectively.

Why are the termination rules not fully attained when the parameter estimates are used? The tailoring algorithm capitalizes on errors in the parameter estimates. As a consequence, tailored testing using the estimated parameters terminates prior to actually reaching the pre-set termination rule. That is, because of capitalization on error in parameter estimates during the process of item selection, the reliability levels implied by the Owen algorithm at any stage during the tailoring process are somewhat inflated. This leads to a too early termination of tailored testing, and, when the obtained $\hat{\theta}$ are correlated with θ, it becomes evident that the pre-set reliability level for termination has not been met. In the present example, an average of 14.51 items was administered when the known parameters were used but only 11.12 when the parameter estimates were used. This shrinkage problem can be overcome by setting the reliability termination rule higher than that actually required. In our present example, the termination rule should be set at .95 in order to obtain $\hat{\theta}$ of reliability .91.

REFERENCES

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished Doctoral Dissertation. University of Washington, 1972.

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.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement,* 1974, 34, 253-269. (*a*)

Urry, V. W. *Computer assisted testing: the calibration and evaluation of the verbal ability bank.* (TS-74-3) Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, December 1974. (*b*)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Manuscript submitted for publication, 1975.

ITEM PARAMETERIZATION PROCEDURES FOR THE FUTURE

FRANK L. SCHMIDT AND VERN W. URRY
U.S. Civil Service Commission

Failure to appreciate the important psychometric role played by guessing in conventional multiple choice tests prevented until recently practical application of latent trait theory to tailored testing. When this problem was properly addressed, it was found that the solution could be expanded to produce an inexpensive and highly accurate item parameterization procedure. Combined with Owen's (1969) elegant Bayesian algorithm and available CRT hardware, these developments made computer-assisted tailored testing feasible from a practical point of view.

The capacity to parameterize new items for possible later inclusion in the item bank during routine operation of the computer-assisted testing system would be a significant step in the direction of even greater practicality (Killcross, 1974). Such a procedure would eliminate the necessity for periodic application of the full parameterization process described by Urry (1975*a*; 1975*b*). The Urry ancillary estimation procedure can be modified to provide the capability to parameterize items in the environment of a live, large-scale, computer-interactive tailored testing system or network. It can thus provide a convenient technology for updating and expanding item banks in ongoing tailored testing systems.

The parameterization procedure is as follows: In addition to the items that are part of his tailored test, each examinee receives a group of additional experimental items. On-line ancillary parameterization can begin for any of these items as soon as a sufficient number of examinees have responded to it. For each item, $\hat{\rho}_{i\theta}$ is computed against the uniformly reliable Bayesian $\hat{\theta}$ from the Owen algorithm. (Notice that the item does not enter in any way into the determination of $\hat{\theta}$.) $\hat{P}_i'$ is estimated in the usual way using sample data. The $\hat{\theta}$ are next grouped into k intervals. Provisional values for $\hat{c}_i$ are assumed, and the minimum χ^2 procedure is applied to obtain approximations of a_i, b_i and c_i. These procedures have been outlined in Urry (1975*b*) and are described in full in Urry (1975*a*).

The purpose of this study was to evaluate the on-line ancillary parameterization process using model sampling and simulation techniques. The one hundred items to be parameterized were those used in the earlier Gugel study, and are shown in Table 1. (In practice, a much smaller number of items would typically be parameterized, but for evaluation purposes a larger number is desirable.) Dependent variables in this study were also the same as those in Gugel's study: correlations between known and estimated parameters and the square root of mean squared deviations of estimated from known parameters. Independent variables are illustrated in Figure 1. Two different banks were used in tailored testing to produce the Owen $\hat{\theta}$, designated as the Verbal Ability Bank and the Ideal Bank. The Verbal Ability Bank of this study consists of the 103 most frequently used items (based on counts from previous simulation studies) from the Commission's 200-item Verbal Ability Bank. The Commission's bank in turn, is made of the best 200 items out of 700 verbal ability items calibrated by Urry (1974). Calibration was carried out on large samples and the final 200 items were chosen to provide a wide distribution of b_i values, high a_i values, and low (below .30) c_i values. The 103 item bank used here thus represents a currently attained—though improvable—level of quality. The Ideal Bank is the same 100 items being parameterized (See Table 1). Three different termination rules were examined for the Ideal Bank; for the Verbal Ability Bank, the most stringent rule (.95) was omitted as impractical. Sample sizes of 1000, 1500, and 2000 were examined. Simulated subjects (θ's) were sampled and their response vectors generated as in the Gugel study. (This procedure is described in full in Urry [1974*a*]).

RESULTS AND DISCUSSION

The obtained standard errors for the Ideal and Verbal Ability banks are shown in Tables 2 and 4, respectively. Tables 3 and 5 present the correlations between actual and estimated item parameters. In most cases, changes associated with variation in the independent variables were in the hypothesized direction. Increasing the number of subjects and the reliabilities required for termination of tailored testing usually resulted in lower standard errors and higher correlations between known and estimated parameters. Some deviation from this pattern occurred because of sampling error. (For each bank, a different sample of simulated subjects was used for each termination rule and sample size examined.) The same is true of the ancillary corrections: the effect was generally to decrease standard errors and increase correlations, but because of sampling error this was not always the case.

In examining the correlations between known and estimated parameters, one should bear in mind that in the case of $\hat{a}_i$, and to a lesser extent $\hat{c}_i$, restriction in range is operating to lower the tabled values. The items parameterized (See Table 2) contained no values of a_i lower than .80. This value of a_i corresponds to a biserial

TABLE 1

True Parameters of the 100 Items Parameterized
Via the On-Line Procedure

Item	Parameters			Item	Parameters		
(i)	a_i	b_i	c_i	*(i)*	a_i	b_i	c_i
1	.80	-1.90	.03	51	1.60	.10	.18
2	.80	-1.70	.06	52	1.60	.30	.21
3	.80	-1.50	.09	53	1.60	.50	.24
4	.80	-1.30	.12	54	1.60	.70	.27
5	.80	-1.10	.15	55	1.60	.90	.03
6	.80	-.90	.18	56	1.60	1.10	.06
7	.80	-.70	.21	57	1.60	1.30	.09
8	.80	-.50	.24	58	1.60	1.50	.12
9	.80	-.30	.27	59	1.60	1.70	.15
10	.80	-.10	.03	60	1.60	1.90	.18
11	.80	.10	.06	61	2.00	-1.90	.21
12	.80	.30	.09	62	2.00	-1.70	.24
13	.80	.50	.12	63	2.00	-1.50	.27
14	.80	.70	.15	64	2.00	-1.30	.03
15	.80	.90	.18	65	2.00	-1.10	.06
16	.80	1.10	.21	66	2.00	-.90	.09
17	.80	1.30	.24	67	2.00	-.70	.12
18	.80	1.50	.27	68	2.00	-.50	.15
19	.80	1.70	.03	69	2.00	-.30	.18
20	.80	1.90	.06	70	2.00	-.10	.21
21	1.20	-1.90	.09	71	2.00	.10	.24
22	1.20	-1.70	.12	72	2.00	.30	.27
23	1.20	-1.50	.15	73	2.00	.50	.03
24	1.20	-1.30	.18	74	2.00	.70	.06
25	1.20	-1.10	.21	75	2.00	.90	.09
26	1.20	-.90	.24	76	2.00	1.10	.12
27	1.20	-.70	.27	77	2.00	1.30	.15
28	1.20	-.50	.03	78	2.00	1.50	.18
29	1.20	-.30	.06	79	2.00	1.70	.21
30	1.20	-.10	.09	80	2.00	1.90	.24
31	1.20	.10	.12	81	2.40	-1.90	.27
32	1.20	.30	.15	82	2.40	-1.70	.03
33	1.20	.50	.18	83	2.40	-1.50	.06
34	1.20	.70	.21	84	2.40	-1.30	.09
35	1.20	.90	.24	85	2.40	-1.10	.12
36	1.20	1.10	.27	86	2.40	-.90	.15
37	1.20	1.30	.28	87	2.40	-.70	.18
38	1.20	1.50	.06	88	2.40	-.50	.21
39	1.20	1.70	.09	89	2.40	-.30	.24
40	1.20	1.90	.12	90	2.40	-.10	.27
41	1.60	-1.90	.15	91	2.40	.10	.03
42	1.60	-1.70	.18	92	2.40	.30	.06
43	1.60	-1.50	.21	93	2.40	.50	.09
44	1.60	-1.30	.24	94	2.40	.70	.12
45	1.60	-1.10	.27	95	2.40	.90	.15
46	1.60	-.90	.03	96	2.40	1.10	.18
47	1.60	-.70	.06	97	2.40	1.30	.21
48	1.60	-.50	.09	98	2.40	1.50	.24
49	1.60	-.30	.12	99	2.40	1.70	.27
50	1.60	-.10	.15	100	2.40	1.90	.03

correlation of .62 between the item and latent ability. Past studies (Jensema, 1972; Urry, 1974) have shown that only about one-third of the items in conventional tests have a_i values this large. No c_i greater than .27 were included; in practice c_i does exceed .27, although the range restriction here is probably not as great as in the case of a_i.

The rather high a_i values among the items parameterized must be considered also in evaluating the root mean square

		Cut-offs*	ITEM BANKS	
			IDEAL BANK	VERBAL ABILITY BANK
SUBJECTS		.91		
	1000	.93		
		.95		
		.91		
	1500	.93		
		.95		
		.91		
	2000	.93		
		.95		

*Reliability values for termination rules.

Figure 1. Experimental Design: Independent Variables

errors for a_i. Errors in $\hat{a}_i$ are much larger for high a_i than low a_i, since when a_i is high, small errors in $\hat{\rho}_{I\theta}$ lead to large errors in $\hat{a}_i$. For example, if $\hat{\rho}_{I\theta} = .90$, $a_i = 2.01$. If $\hat{\rho}_{I\theta} = .88$, $\hat{a}_i = 1.85$, a difference of .16. But if $\hat{\rho}_{I\theta} = .50$, $a_i = .58$. Then if $\hat{\rho}_{I\theta} = .48$, $\hat{a}_i = .55$, a difference of only .03.

The real test of the usefulness of the on-line parameterization process lies in the performance of the parameter estimates in tailored testing. The better the estimates, the closer they will come to equaling the performance of the known parameters. The parameter estimates obtained in this study have not yet been used in simulated tailored testing, but an idea of how well they would perform can be obtained by examining the performance of parameter estimates from Gugel et al. (1975) with roughly equivalent errors. Table 6 compares root mean square errors and correlations between known and estimated parameters from the present study for the Verbal Ability Bank with 2000 cases and reliability cut-off of .93 with the results obtained by Gugel et al. (1975) using 1000 cases and 60 items with the full parameterization process. Except for the standard error of $\hat{b}$ (which is lower) and $r_{\hat{a}a}$ (which is also lower), his results are essentially equivalent. Using a reliability cut-off of .95, Gugel et al. conducted simulated tailored testing using both the known and the estimated parameters. Known parameters produced $r_{\hat{\theta}\theta} = .9752$, exactly corresponding to the termination rule (i.e., $[.9752]^2 = .95$). With the parameter estimates, $r_{\hat{\theta}\theta}$ was .9516, corresponding to an obtained reliability of .9044.

Because the tailoring algorithm capitalizes on chance errors in the parameter estimates, tailored testing using the estimated parameters is terminated prior to actually reaching the pre-set termination rule. That is, because of capitalization on error in parameter estimates during the process of item selection, the reliability levels computed by the Owen algorithm at any stage during the tailoring process are somewhat inflated. This leads to a too early termination of tailored testing, and, when the obtained θ are correlated with θ, it becomes evident that the pre-set reliability level for termination has not been met. In the present example, an average of 14.57 items was administered when the known parameters were used but only 11.12 when the parameter estimates were used. This shrinkage problem can be overcome by setting the reliability termination rule higher than that actually required. In our present example, the termination rule should be set at .95 in order to obtain $\hat{\theta}$ of reliability .90. Simulation studies provide a convenient—and perhaps the only—method of determining in advance of actual use the amount of shrinkage to be expected when items are parameterized on given sample sizes and with given numbers of items. The shrinkage problem here is thus somewhat different from that characterizing, say, multiple regression, in that its effects can be cancelled out by appropriate selection of termination rules. Two points, however, should be noted here:

1. Parameterizing on large sample sizes (both numbers of items and numbers of cases), and thus obtaining more accurate initial parameter estimates, is preferable where feasible to adjusting termination rules to allow for shrinkage.
2. For certain tailored testing usages—for example, battery tailoring or multivariate tailored testing—the advantages of parameter estimates that can fully meet pre-set termination rules become substantial. That is, adjustment of termination rules to allow for shrinkage becomes, at best, inconvenient and awkward.

In light of these facts, an important question is whether or not the on-line parameterization process can produce estimates with errors low enough to reduce shrinkage to negligible levels. An important consideration, of course, is the quality of the item bank on which the original $\hat{\theta}$ are derived. By parameterizing and adding to the Verbal Ability Bank those items which were erroneously rejected earlier on the basis of low point-biserial and biserial item-total indices, it will probably be possible to make the Verbal Ability Bank equivalent to the Ideal Bank used in this study. By increasing the number of cases to 3000, or perhaps beyond 3000, it should be possible to reduce the

TABLE 2

Root Mean Square* For Item Parameter Estimates And $\hat{\rho}_{I\theta}$ Using the Ideal Bank

Subject	Cut-offs	Uncorrected				Corrected			
		a_i	b_i	c_i	$\rho_{I\theta}$	a_i	b_i	c_i	$\rho_{I\theta}$
1000	.91	.465	.226	.089	.076	.340	.174	.086	.057
	.93	.480	.227	.095	.075	.357	.164	.079	.054
	.95	.418	.202	.093	.068	.283	.187	.074	.045
1500	.91	.481	.189	.086	.079	.318	.225	.075	.051
	.93	.467	.202	.091	.079	.290	.208	.067	.049
	.95	.445	.193	.095	.071	.311	.206	.070	.047
2000	.91	.506	.232	.091	.082	.267	.236	.079	.044
	.93	.477	.218	.090	.071	.270	.198	.067	.042
	.95	.454	.209	.090	.071	.297	.203	.066	.042

*RMSE = $\left(\frac{\Sigma(p_i - \hat{p}_i)^2}{n}\right)^{1/2}$ where p = parameters
n = number of items

TABLE 3

Correlations Between Known and Estimated Parameters—Ideal Bank

Subject	Cut-offs	*Uncorrected*			Corrected		
		a_i	b_i	c_i	a_i	b_i	c_i
1000	.91	.807	.995	.567	.820	.994	.548
	.93	.780	.994	.495	.780	.994	.540
	.95	.876	.994	.504	.874	.995	.553
1500	.91	.844	.996	.617	.832	.995	.656
	.93	.861	.995	.593	.860	.995	.624
	.95	.857	.995	.567	.852	.995	.610
2000	.91	.883	.995	.610	.886	.995	.631
	.93	.892	.996	.602	.892	.996	.641
	.95	.883	.996	.617	.883	.997	.649

TABLE 4

Root Mean Square Errors* For Item Parameter Estimates And $\hat{\rho}_{I\theta}$ Using the Verbal Ability Bank

Subject	Cut-offs	Uncorrected				Corrected			
		a_i	b_i	c_i	$\rho_{I\theta}$	a_i	b_i	c_i	$\rho_{I\theta}$
1000	.91	.596	.259	.093	.103	.370	.261	.097	.055
	.93	.599	.285	.093	.107	.400	.258	.095	.060
1500	.91	.514	.208	.090	.081	.280	.267	.084	.048
	.93	.554	.286	.082	.098	.336	.267	.075	.050
2000	.91	.562	.217	.087	.096	.338	.275	.076	.043
	.93	.553	.257	.086	.096	.331	.250	.072	.045

*RMSE = $\left(\frac{\Sigma(p_i - \hat{p}_i)^2}{n}\right)^{1/2}$ where p = parameter,
n = number of items.

TABLE 5

Correlations Between Known And Estimated Parameters–Verbal Ability Bank

		Uncorrected			Corrected		
Subject	Cut-offs	a_i	b_i	c_i	a_i	b_i	c_i
	.91	.786	.993	.524	.780	.933	.550
1000	.93	.821	.993	.510	.807	.993	.515
	.91	.875	.994	.565	.875	.994	.594
1500	.93	.871	.993	.614	.870	.993	.624
	.91	.836	.996	.622	.819	.995	.655
2000	.93	.878	.996	.562	.879	.996	.591

TABLE 6

Comparison of Gugel Results with Present Study Results

	Root Mean Square Errors				Correlations $(r_{\hat{p}p})$		
	a_i	b_i	c_i	$\rho_{I\theta}$	$r_{\hat{a}_i a_i}$	$r_{\hat{b}_i b_i}$	$r_{\hat{c}_i c_i}$
Gugel (1975)*	.322	.140	.062	.044	.842	.995	.588
Present Study**	.331	.250	.072	.045	.879	.996	.591

*N = 1000, 60 items; full parameterization procedure.
**Verbal Ability Bank, N = 2000, Relaibility cut-off = .93.

root mean square errors shown in Table 2 (2000 cases, cut off at .95) to levels comparable to those obtained by Urry (1975) with the full parameterization process (2000 cases, 100 items). Urry's root mean square errors were .242, .123, and .056 for $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$, respectively. At this level of accuracy, little shrinkage was in evidence. It should be borne in mind that, in the case of the on-line parameterization process, the number of cases can be increased at little or no cost. Also, as the quality of the bank is increases, more stringent termination rules can be introduced, further increasing accuracy of the on-line parameter estimates.

A final modification of the on-line parameterization process can be made which should further reduce estimation errors. As the parameterization procedure is presently set up, those examinees whose $\hat{\theta}$ do not attain the termination rule reliability within 30 items are dropped from the sample. Because coverage of θ is weakest in the Verbal Ability Bank in the low ranges, the dropped subjects tend to be concentrated in the low end of the distribution. This creates a paucity of information in a range in which many c_i values are determined, leading to higher c_i errors. Also, when the truncated distribution is restandardized, the result is a displacement of the $\hat{b}_i$ values. In the case of the Ideal Bank, no subjects were dropped at the .91 and .93 termination rules. Even at the .95 termination rule few examinees failed to reach the criterion (10 at N = 1000, 8 at N = 1500, and 9 at N = 2000). In the Verbal Ability Bank, no subjects were dropped at .91, but at .93, 23 were dropped at N = 1000, 53 at N = 1500, and 40 at N = 2000. Thus, up to 3.5% were eliminated. This probably explains to a great extent the failure of the .93 termination rule to produce noticeably better estimates than the .91 rule (Tables 4 and 5). Estimates would probably be improved by retaining in the sample those subjects who fail to reach the termination rule within 30 items. Although these $\hat{\theta}$ are less reliable, they probably provide information at low θ which is useful for parameterization purposes.

REFERENCES

Gugel, J., Schmidt, F. L., & Urry, V. W. *Effectiveness of the ancillary estimation procedure.* Paper presented at the Conference on Computerized Adaptive Testing, Washington, D.C., June 1975.

Jensema, C. J. *An application of latent trait mental test theory to the Washington pre-college test battery.* Unpublished doctoral dissertation, University of Washington, 1972.

Killcross, M. C. *A tailored testing system for selection and allocation in the British Army.* Paper presented at the 18th International Congress of Applied Psychology, Montreal, August 1974.

Owen, R. J. A Bayesian approach to tailored testing. *Research Bulletin,* 69-92. Princeton, N.J.: Educational Testing Service, 1969.

Urry, V. W. *Ancillary estimators for the item parameters of mental test models.* Personnel Research and Development Center, U.S. Civil Service Commission, 1975 in press (*a*).

Urry, V. W. *A five year quest: is computerized adaptive testing feasible?*Paper presented at the Conference on Computerized Adaptive Testing, Washington, D.C., June 1975 (*b*).

Urry, V. W. *Computer-assisted testing: calibration and evaluation of the verbal ability bank* (TS 74-3). Washington, D.C.: Personnel Research and Development Center, U.S. Civil Service Commission, December 1974.

DR. FREDERIC M. LORD
Educational Testing Service:

It is appropriate that my discussion should be expressed in the first person singular—to continually remind you that I am giving my own opinions, which may be biased, since I am not a disinterested party here. There have been many, many important points made during these sessions. I have chosen 14 points to emphasize in my discussion.

1. Cliff (Note 1) writes: "It is felt that our formulation will provide the framework for a test theory which is more appropriate to the interactive case than either the classical or traceline theories are." I am sure he would not want this challenge to ICC theory to go unanswered. Cliff proposes that the appropriate model for the item responses is the Guttman scale.

Since the Guttman scale is a special case of the more general logistic or normal ogive item characteristic curve, I cannot see how the Guttman scale can be called a more appropriate model than the logistic or normal ogive. If the Guttman scale were the correct model, the fitted logistic or normal trace lines would come out in the Guttman form.

The Guttman scale assumes that the tetrachoric correlation between any two items is 1.00. This value may be approximated for certain attitude test data, but for aptitude and achievement test data, typical tetrachoric item intercorrelations are usually less than 0.35. This is so *very* different from 1.00 that I cannot see how the Guttman model can be considered acceptable for aptitude and achievement tests.

2. Consider the problem of testing and assigning new armed forces recruits. One recruit, perhaps, should take a complete battery of tests to determine his suitability for officer training school. The next recruit, however, should be quickly extricated from this battery of tests and perhaps given a battery of mechanical aptitude tests. How can we use adaptive testing to route a new recruit through many such batteries of tests efficiently, with a minimum waste of time? Glenn Bryan raised this important question with me some years ago. It seems as if adaptive testing should be an excellent way to deal with this problem. Yet the situation is so multidimensional that current theory does not tell us how to proceed. Here is a very important unsolved problem.

3. Waters has pointed out and documented something that some of us had overlooked--that an adaptive test should be expected to take longer to administer than a conventional test with the same number of items. The reason is that the conventional test contains items that are too hard or too easy for each examinee—items that he can answer (or omit) without need for lengthy consideration. Studies of adaptive testing will have to take testing time into account.

4. There is one situation in which adaptive testing (or some other unconventional procedure) is really indispensable. Suppose it is necessary to have good measurement over an unusually wide range of ability. As a first step, one might build a conventional type of test with extra easy items added at one end and extra hard items at the other, so as to have some items that are appropriate in difficulty for each ability level. Of course, the easy items are a waste of time for the high-level examinees, but that is not the serious problem. The hard items are not merely a waste of time for the low-level examinees. The guessing of low-level examinees on the hard items adds so much noise that the measurement provided by the easy items is nearly drowned in random error.

In such situations, it can be shown that the test would be much improved as a measuring instrument for low-level examinees if we simply threw away (or refused to score) the more difficult half or two-thirds of the test. The situation cannot be remedied simply by adding more easy items. If we wish to obtain good measurement at low as well as at high ability levels, some kind of tailoring is necessary so that hard items are not administered to low-level examinees.

5. If total testing time is held fixed, adaptive testing leads to better measurement for some examinees. If accuracy of measurement is held fixed, adaptive testing leads to reduced testing time for some examinees. These two alternatives are not basically different.

Keeping the standard error of measurement fixed across examinees would be simple if the test were very long or if we knew the true parameter values, and if all items had identical characteristic curves. Otherwise there may be difficulty in finding a good small-sample theory and method. Gugel and Schmidt have given empirical evidence of this. This is a problem in sequential estimation (Wald, 1951; Robbins, 1959; Bickel & Yahav, 1968). Except perhaps for Bayesians, methods of sequential estimation are not as well settled as are methods of sequential hypothesis testing. Even sequential hypothesis testing poses unsolved problems when the items do not all have identical characteristic curves.

6. It is undoubtedly significant that most of the speakers here are using two- or three-parameter item characteristic curve models. No one here has urged that adaptive testing be limited to the one-parameter Rasch model.

It is sometimes asserted that the Rasch model is the only one that allows us to estimate examinee ability independently of the items administered. I would argue that all ICC models allow us to do this. The unique virtue of the Rasch model is that it provides a sufficient statistic for estimating examinee ability. Sufficient statistics are desirable, but they are not common in statistical work, outside of the usual normal-curve theory. Statistical inference still proceeds very effectively in the absence of sufficient statistics.

The objection usually cited against the Rasch model is that it assumes all items to be of equal discriminating power. I suspect that an even more serious objection is that it assumes there is no guessing. Any attempt to modify the Rasch model to take guessing into account would necessarily destroy the sufficiency properties of the Rasch model that make it attractive.

7. This brings us face to face with the question whethei to use a two- or a three-parameter ICC model. Waters used a two-parameter normal-ogive model and the assumption that ability is normally distributed to estimate the a parameters (discriminating power) of the 50 verbal items in Form 2B of SCAT II. By chance, I had available estimates of the same parameters based on the three-parameter logistic model, computed by a program called LOGIST (available on request).

I have plotted Waters' values against the LOGIST values in Figure 1. Each point is shown as a digit representing item difficulty. The larger the digit, the more difficult the item and the more the examinees' responses are affected by guessing. Agreement is good only for the easy items where there is no guessing.

Many studies comparing different estimation methods should be carried out. Some should use real data; some should use artificial data, where the true parameters are known. I should be glad to run on LOGIST any suitable set of data that someone here may wish to use for making such comparisons.

8. In the three-parameter models, the ICC's have the form $c_i + (1 - c_i)F[a_i(\theta - b_i)]$. This mathematical form is not beyond challenge, as Samejima has pointed out, but it is relatively easy to defend as a versatile form that fits the data, so long as we do not suggest that examinees either know the answer to the item or else guess with probability of success c_i. We all know that examinees do not respond this way. If ICC theory were based on the dichotomy, knowledge or random guessing, it would not be credible. For this reason, it may be best not to refer to c_i as a 'guessing parameter.' (I confess to violating this good advice.)

9. When working with real answer sheets, it becomes necessary to deal with the problem of omitted responses. If we require the examinee to answer all items, we are purposely introducing random error into our data. In addition, we are forcing an examinee who has demonstrated a certain level of performance by his responses to gamble on some possibily random events, which may, if he is unlucky, destroy all the positive evidence of ability that he has displayed.

If we permit the examinee to omit items, we cannot properly treat such responses as wrong. To do so would penalize the examinee who omits, in comparison to the examinee who guesses.

It seems at first thought that we might simply treat omitted items as if they had not been administered at all. This cannot be correct, however. If we ignore omitted items, an examinee could win a very high estimate of ability simply by answering items only when he was completely sure of his answer.

The fact that an examinee has omitted an item carries information about his level that cannot be ignored. A method for using this information efficiently, under certain assumptions, is outlined in a *Psychometrika* paper (Lord, 1974).

10. I want to take this opportunity to make a correction. In a 1968 paper (Lord, 1970), I wrote:

> If $a_i = 0.333$, under the assumptions already made [the] reliability for a 60-item test will be 0.80; if $a_i = 0.5$, this reliability will be 0.90; if $a_i = 1.0$, this reliability will be 0.97. In view of this, we shall choose $a_i = 0.5$ as a typical value and shall address most of our attention to it.

After seven years of experience with the a parameter, these reliabilities sound high. Actually, they are correct, but, as the assumptions stated, they are for free response, not multiple-choice items. Urry made this same point this morning. Since most of the cited paper dealt with multiple-choice items, it was a mistake to suggest $a_i = .50$ as a typical value. Although the diagrams presented in that paper required the reader to supply his own values of a_i, the general impression given was one of only limited enthusiasm for adaptive testing.

Current results show that when $a_i = 0.9$, a peaked test composed of 40 five-choice items should have a KR_{20} reliability of .90. When a_i is 0.9, the conclusions supplied by the diagrams in the cited paper are quite encouraging for the future of adaptive testing.

11. The purpose of the cited paper was to evaluate adaptive tests in comparison to conventional tests. To do this, the situation considered had to be a simple one. This was the reason for the use of a fixed-step-size up-and-down branching procedure. Such a procedure is *not* to be recommended for practical testing.

When the item parameters have been estimated and a computer is available for making the calculations, the choice of the item to be administered next should be made by checking all unused items (perhaps within a specified item type) and selecting the item that is expected to give the most information about the examinee.

If a Bayesian prior distribution of ability is being used, and if this distribution is normal, this is Owen's (in press) procedure, frequently used today. In such a procedure, except for certain approximations each step is locally optimal. We cannot expect local optimality to produce overall global optimality, but the difference may not be of great importance.

12. When we select the next item to be administered on other considerations besides item difficulty, we no longer

have an up-and-down branching procedure. The next item administered after a correct response might be an easier item, not a harder item.

The recommended procedure means that items with high a_i will be used very frequently and items with low a_i will be used seldom or not at all. The gain from this use of the best items will probably more than double the gain from any procedure, such as the up-and-down procedure, that selects items solely on item difficulty.

Furthermore, the larger the item pool, the greater the gain. This is not surprising. We always knew that if we

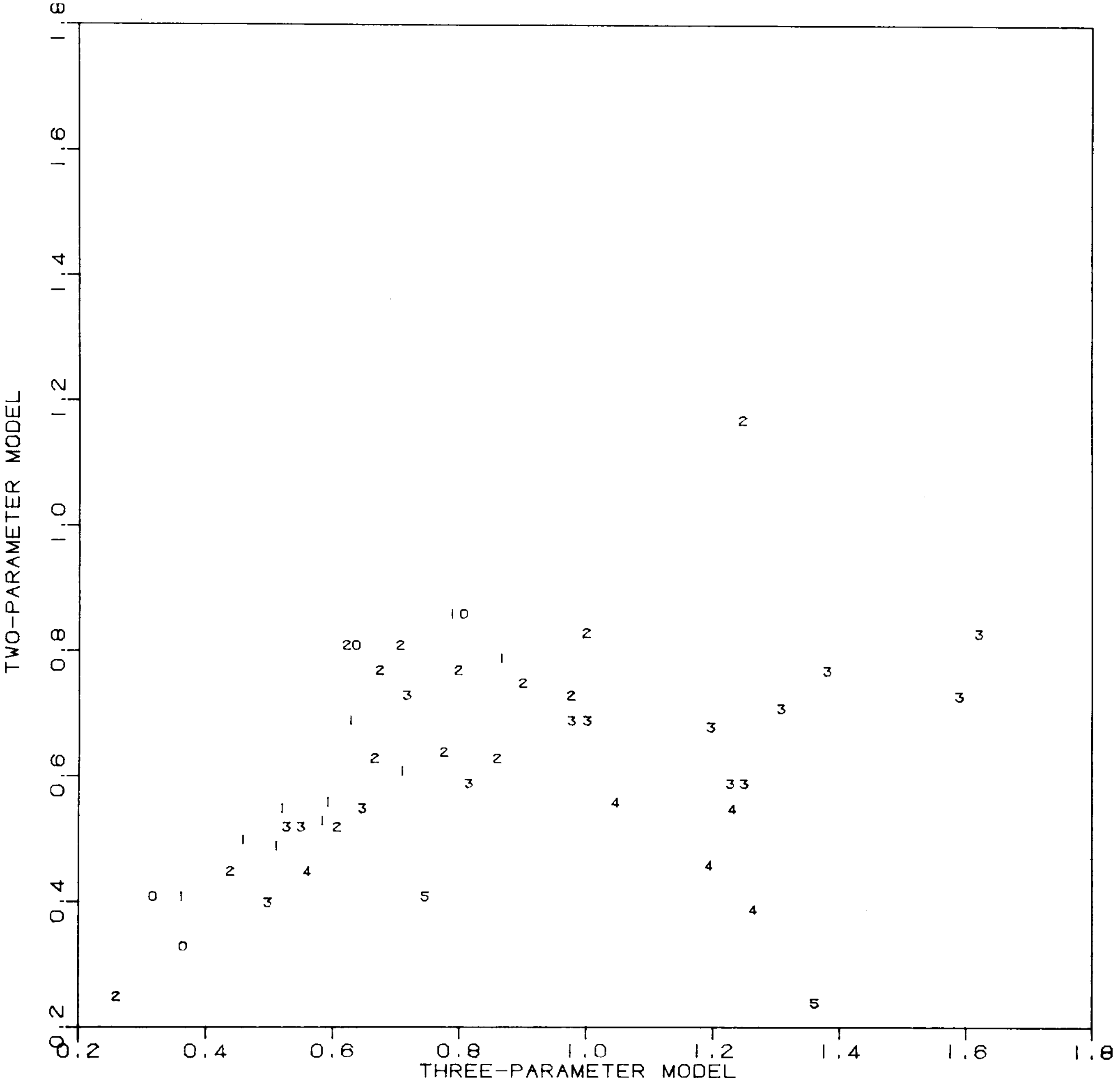


Figure 1. SCAT 2B. A comparison of estimated a_i parameters. The two-parameter model assumes a normal distribution of ability. Each item in the plot is located by a digit which represents item difficulty (b_i + 3). The easiest items are indicated by a 0, the hardest by a 5.

selected the best items from ten tests, we could build a single test that would be much more reliable than any of the original tests.

13. My last point concerns the use of Bayesian inference in adaptive testing. When we are testing large numbers of examinees all coming from a single source, we are in a really exceptionally good position to obtain and use a prior distribution describing the examinees. It would seem negligent not to obtain and use such a readily available prior distribution.

On the other hand, I would like to make a simple point not often expressed. Bayesian inference based on a prior distribution will give correct results when the prior corresponds, in some sense, to reality. It is likely to give incorrect results if the prior itself is incorrect.

In most Bayesian work, it is usually not practicable to determine whether the prior is correct or incorrect. In our work, on the contrary, it is fairly easy to do so. We need only estimate the ability of each person tested and then look at the distribution of estimated abilities.

If we were testing unselected school children in grade school, a normal distribution of ability might possibly be found. When we are testing highly selected groups in college or elsewhere, it seems unlikely that we will find a normal distribution.

Bayesians point out that the effect of an assumed prior becomes unimportant as the number of observations becomes large. In our context, this means that our ability estimates will not be spoiled by an incorrect prior distribution of ability provided the test administered is long enough.

This is not the whole story, however. The assumption of a normal distribution of ability, if false, may lead to unsatisfactory estimates of item parameters. The usual formula for biserial r can give absurd results if the continuous variable, in this case examinee ability, unknown to the statistician, is far from normally distributed. Unlike some other effects of Bayesian priors, this difficulty does not diminish as sample size becomes large.

Two different estimates of the distribution of examinee ability for one set of data are shown in Figure 2, reproduced here from Lord (1974). The agreement between the two estimates, obtained from very different assumptions, gives me some confidence in these results. My empirical results from other sets of data (including a representative sixth-grade group) are similar. When the ability scale is chosen so that all item characteristic curves are three-parameter normal ogives, or logistic curves, it turns out, for my data, that ability is not normally distributed.

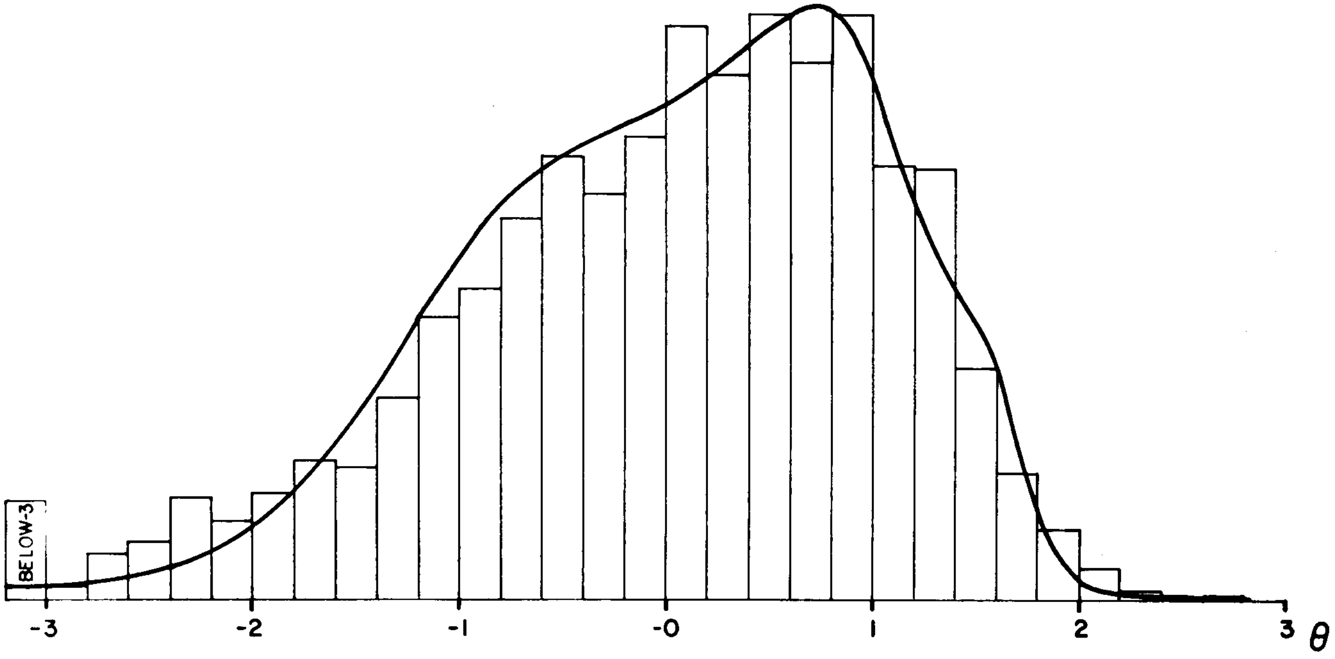


Figure 2. Distribution of estimated θ (histogram) and estimated distribution of θ (curve). Reproduced from Lord (1974) with permission of *Psychometrika.*

14. Although I an not a market analyst, I will without much risk venture two assertions. Computer costs—if they have not already done so—will come down to the point where computer-based adaptive testing is economical. When this happens, adaptive testing will come into wide use. The

McKillip and Urry paper provides important details on this subject.

REFERENCE NOTE

1. Cliff, N. *Complete orders from incomplete data: Interactive ordering and tailored testing.* Mimeographed paper. Conference on Computerized Adaptive Testing, Washington, D.C., June 1975.

REFERENCES

Bickel, P. J., & Yahav, J. A. Asymptotically optimal Bayes and minimax procedures in sequential estimation. *The Annals of Mathematical Statistics,* 1968, *39,* 442-456.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance.* New York: Harper and Row, 1970.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika,* 1974, *39,* 247-264.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, in press.

Robbins, H. Sequential estimation of the mean of a normal population. In U. Grenander (Ed.), *Probability and statistics.* New York: Wiley, 1959.

Wald, A. Asymptotic minimax solutions of sequential point estimation problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1951.

DR. BERT F. GREEN, JR.
Johns Hopkins University:[1]

Tailored testing has been talked about for many years in academic circles. In this conference we have heard firm plans for action. The promise of tailored testing is becoming real. Numberless simulated examinees have taken tailored tests and a substantial, though smaller, number of real people have also had the experience. The use of tailored tests will provide substantially improved efficiency and will have a number of beneficial side effects, as mentioned by McKillip and Weiss, among others. Testing conditions will be more nearly standardized, the test will hold the taker's interest because each item will be a challenge, possibly there will be less test anxiety, feedback may decrease racial bias. (Weiss; Johnson, 1973)[2]

There will also be some harmful side effects, that we may as well face. People will have trouble understanding the system, and complaints will be frequent. Two people with widely different abilities will both experience getting about half the items rights, yet get very different scores; one will be accepted, the other rejected. If these two people compare notes, they may be confused. The anti-testing forces are also for the most part anti-computer, so negative voices will be raised. Security is at least as difficult with a computer system as with a paper and pencil system. But these are operational problems, and now is not the time to worry about them. They will all be solved, somehow. I merely list them to counter the tendency to believe that the millenium is upon us.

Now let me make one thing perfectly clear. I am about to criticize aspects of the work reported at this conference. That is my job. But the one most important fact, that outweighs all criticism, is this: The operational use of tailored testing is a giant step forward in personnel evaluation. Evidence indicates as much as a 2 to 1 gain in efficiency, and possibly some very important side benefits. I am completely convinced that this is an important step to take. My comments are of two kinds—suggestions for clarifying and improving the theoretical basis for this big step, and impatience at our not yet having planned further giant steps. These steps should be justified not in terms of saving money, which Hansen claims, but in terms of doing a better job.

Let us now consider some of the technical problems in a computer-based system. We have heard two plans for item analysis "on-the-fly", as they say in the computer trade. A question arises about some of the item analysis procedures (Urry; Jensema) which still seem to be built on the biserial correlation of the item with the ability scale, and the overall proportion of correct answers. These raw data are reparameterized (to use an ugly word that should be banned from civilized discourse) but the basic data are $\rho_{\theta i}$ and P_i. Both of these indices depend on the notion of a population of test takers. Yet one purpose of tailored testing is to avoid the notion of population. What, for example, is the population for Lord's broad-range flexilevel test of verbal ability? Everyone from fifth grade to college? In tailored testing, it would seem that the item parameters must be based on the regression of the item on the ability scale. This sounds a little circular—perhaps it is. Some sort of iterative optimization process would be needed at the start, to ever get the ability scale in the first place. Cliff described one such procedure for his ordinal scale model; an equivalent procedure could easily be devised for the metric model.

Cliff's procedure also depends on a population. He goes so far as to say that the purpose of a test is to rank order the population of examinees. Sometimes it is, but often it is not. Often the purpose is to categorize the examinee as qualified or not qualified for a particular job. Or even better, to give a quantitative index of the degree of qualification. The only population we are really interested in is the population of successful job holders.

There are other technical problems with Cliff's scheme, which he promises to solve. For example, he did not describe what happens when a person's item responses have contradictory implications for other cells in his matrix. Indeed his system probably tries to avoid asking questions that might provide contradictory information.

The main reservation I have about the technical side of tailored testing is the commitment to latent trait theory. The concept of a latent ability scale is a great improvement over the concept of a true score. The true score model was never a very good idea; rather, it was a simple model that worked pretty well. But are we sure that the latent ability score is much better? Does the latent trait model fit the tests for which it is used? Is the assumption of local independence really tenable? Suppose, for example, that there are secondary factors in common among subsets of items. How much difference would that make? Nobody knows.

The point is that latent trait theory *is* a theory, just as any other behavioral theory, and it needs verification. Empirical work is needed to show that latent ability scores work as the theory predicts. Simulated examinees will not do—studies are needed with real people. Are the scores invariant over item selections, or over samples of individuals? Does the precision of measurement really work the way the information variable says it does? What about the relation of validity to test length or information? Empirical

[1] This work was done with support from Grant GB37520 from the National Science Foundation. The author is indebted to Warren S. Torgerson for many fruitful discussions of computer applications in testing and personnel decision.

[2] Throughout, references to other papers in this conference are by author only; other references are followed by publication year.

work has been presented by Waters and others, but whether it supports the theory is not clear.

Classical test theory has a curious status: most psychologists and educators believe that it is fact, not theory. Nowhere in Lord & Novick's treatise is there a section on empirical verification of the theory. Actually, test theory is a self-consistent, much-elaborated theory that seems to work pretty well. For example, the Spearman-Brown formula usually works. Some people look upon the Spearman-Brown formula as a fact. It is a fact only in the sense that it is a logical consequence of the basic assumptions of the theory. So far as I know, neither true score theory nor latent trait theory has been put to a critical test, as have most other mathematical theories of behavior.

One final theoretical issue needs clarification. The literature contains results (e.g., Lord, 1970) indicating that a tailored test is not much more effective than an ordinary test with a peaked item difficulty distribution. The advantage lies mainly in the extremes. But the theoretical and empirical results presented in this conference indicate that a tailored test is much better even in the mid-range. Work is needed to clarify when a tailored test will help and when it won't.

One final point about technical terminology. In the simulation studies of Jensema, Waters, McBride, and others, the estimated ability $\hat{\theta}$, which is the test score in tailored testing, is supposed to be nearly θ. The closeness of $\hat{\theta}$ to θ is measured both by $(\Sigma(\theta - \hat{\theta})^2/N)^{1/2}$, which was called the "standard error" and by $r_{\theta\hat{\theta}}$, which was called the "validity". In engineering, the former measure is commonly called the root-mean-square error, or R.M.S. error; it is not, after all, a standard error, since it's not a standard deviation. Mean square error includes both error variance and squared bias. Thus the measure is very appropriate; but it is misnamed. To call $r_{\theta\hat{\theta}}$ the "validity" is much worse; it is downright sinful. This use of the term goes back, I'm told, to Ledyard Tucker and Hubert Brogden, but that only proves that people in high places make mistakes. A different word must be used. "Validity" is seriously misleading, and has even been mis-interpreted at this conference. My own candidate for a name for $r_{\theta\hat{\theta}}$ is "fidelity". I hope the in-group either uses "fidelity" or finds another word.

Next Steps

Now that tailored testing is about to become operational, perhaps it is time to take a longer-range perspective. Do the present developments really exploit the power of an interactive computer? Many scientists, in their first encounter with a computer, use the computer mainly to do faster and neater what they were already doing before computers. It is as if the horse and buggy industry's reaction to internal combustion engines had been to build a mechanical horse. Statistical computation is a good case in point. To a very large extent, statistics is still at the mechanical horse stage in its use of computers. The statistical program packages are fast ways to do old things—analysis of variance, regression, factor analysis. Even the few new things, such as nonmetric scaling and clustering, had their roots in pre-computer ideas. Interactive statistical methods are still in their infancy. Mostly, interaction means replacing the control cards in an input deck by questions printed by the machine and answered by the user on the spot. No subtle interplay of human judgment and computer speed is implied.

The mechanical horse stage in computerized testing would be an automatic test production system. Given the characteristic of a population, the computer would select the most appropriate items from its item files and would print a suitable test. I naively thought testing had avoided this typical first stage, but apparently such systems were built, some years ago.

Tailored testing is one step beyond the mechanical horse stage. To be sure, the up-and-down method had seldom been used in mental testing, barring Binet, who didn't do it right, but the up-and-down method is an old stand-by in psychophysics, and in sensitivity testing generally, dating from World War II and earlier. Also, test theoreticians knew that measurement was best when the items were all sufficiently difficult that the examinee got about half of them correct. (Actually about 68% for 5-alternative items, Fred Lord reminds me, because of guessing.) This is one part of the theory that none of the operational people believed, but the theory was there. So the adaptive test was a natural next step in computer involvement in testing. Still, the only use of the computer in tailored testing, apart from the trivial use in presenting the items on a terminal, is in selecting the next item and computing the ability score. The same 5-choice items are being used, the item is scored either right or wrong, the same kinds of traits are being measured. Now is the time to move on, in research at any rate, to better things.

Many more opportunities exist. Some have been mentioned at this conference. Samejima proposes that we use the particular wrong choice of an item as partial information. Some wrong choices are better than others. Item response weighting has minimal utility in standard tests, primarily because of the test length. Weighting becomes more useful with fewer items, which is just what tailored testing provides. In addition to Samejima's proposal, even more information could be obtained, when the response is wrong, by asking for a second try. The procedure of trying alternatives until getting the right answer goes back to the 1940's or earlier. In those days, Science Research Associates sold a punch board on which answers were punched out. Instructions were to punch out alternatives until the red dot appeared, signalling the right choice. The item score was the number of unpunched choices, except that omits got a negative score. I am told that test scores based on these item scores were consistently more reliable and more valid than scores based on a 1-0 item scoring. The computer terminal is an elegant punch-board! Another possibility is to have the examinee rank or rate the alternatives for suitability. The probability assignment proposal of Shuford et. al., (1966) now being tried by Weiss and his coworkers is equivalent; though the restriction that the ratings must add to one, like probabilities, is an unfortunate complication that is likely to have adverse operational consequences. Ratings or rankings would be better.

The computer permits the use of constructed responses—fill in the blanks—rather than multiple choice. Computer processing of constructed responses has been worked on in computer assisted instruction; these techniques could be adapted to the testing situation. Most of our present item types have evolved in a multiple choice environment, and constructed responses would be no help. For example, some verbal analogies items would not work as constructed responses – e.g., "Brick is to building as leather is to________." Others would work: "Shoe is to foot as helmet is to________." The difficulty of vocabulary items is controlled almost entirely by the distractors, so asking the examinee to construct a synonym would markedly alter the item. But there is no reason why new item types cannot evolve in the new context. Verbal fluency is a natural for the computer to test, and virtually impossible in the multiple choice context.

Of even more interest is the possibility of new types of items, and new types of traits. The GRIP tests of Cory are especially interesting, as are some of the items briefly mentioned by Weiss, such as his conceptual maze. Many of these types can be tried on present day alphanumeric terminals, others need graphic terminals, which are at present too costly, but which may soon be relatively inexpensive.

I am convinced that the potential for new styles of items, or contingent sets of items, is the next important contribution of the computer. After all, we already know how to measure verbal ability and quantitative ability. The computer merely gives us efficiency. What we need is more information.

The computer could also be immensely helpful if we placed less emphasis on measurement and more on the decision process. Instead of providing a test battery, we could provide a decision system. Many years ago Cronbach & Gleser (1965) argued for the necessity of coupling the decision process with the testing process. The computer, and computer assisted testing, have provided an unparalleled opportunity to do this. Hansen, McKillip, & Lord have mentioned this.

Consider the simple example of selecting among applicants for a particular job or for entry to a particular college. The test's job is to label each taker as qualified or not qualified. This implies a cut-off score, or at least a cut-off region. The very well qualified and the very poorly qualified persons can probably be identified relatively quickly; most of the effort should be spent on the borderline cases. To be sure, we must beware of Lord's lucky guesser, and Weiss' low consistency scorer, but with care, an efficient system can be devised that does not measure accurately at all levels, but only where it counts.

A one-dimensional case is only the beginning. Both Weiss and Hansen have suggested that additional savings can be made when there are several relevant dimensions. Here, progress requires that the decision process be coupled with the testing process to build a complete system.

There are many different approaches to a personnel decision system. One model would treat jobs as regions in a space whose dimensions are specific job requirements, specific abilities, or characteristics needed for the job. A person is a point in this space, the testing problem is to pinpoint the person's position sufficiently accurately to be able to list the jobs for which he is qualified, and possibly to list these in rank order from the ones for which he is most qualified to the ones for which he is barely qualified. The dimensions of the job space might be abilities, or they might not. And individual items might serve to locate a person on only one dimension, or items might help to locate a person in the total space. At least, there is no *a priori* reason for discarding impure multidimensional items. Indeed such items might be especially useful in a decision system.

Five years ago at a similar conference (Green, 1970) I said that the computer had a great future in testing. Today, happily, it has a present as well as a future. Operational versions of tailored tests represent a great technical achievement. Furthermore, the computer plays a central role in the enterprise. Still, the potential of the computer has barely been tapped. The future lies ahead.

REFERENCES

Cronbach, L. & Gleser, G. C. *Psychological tests and personnel decisions.* 2nd ed. Urbana, Illinois: University of Illinois Press, 1965.

Green, B. F. Comments on tailored testing. In Holtzman, W. H. (ed.) *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Johnson, D. F. and Mihal, W. L. Performance of blacks and whites in computerized vs. manual testing environment. *American Psychologist,* 1973, *28,* 694-699.

Lord, F. M. Some test theory for tailored testing. In Holtzman, W. H. (ed.) *Computer-assisted instruction, testing, and guidance.* New York: Harper & Row, 1970.

Shuford, E. H., Albert, A., and Massengill, H. E. Admissable probability measurements procedures. *Psychometrika,* 1966, *31,* 125-145.

ANNOUNCEMENTS

Dr. Robert J. Gettelfinger of Educational Testing Service announced that organization's willingness to edit a newsletter on the subject of computer-assisted testing. He asked for suggestions as to the content of the newsletter, and for the opinions of the conferees as to what subject matter should be covered and as to whether contributions should be entirely voluntary or should be obtained by assigning papers.

Dr. David J. Weiss of the University of Minnesota announced that he will edit a new journal, *Applied Psychological Measurement*, that will publish empirical research on the application of techniques of psychological measurement to substantive problems in all areas of psychology and related disciplines such as sociology and political science. He invited conference participants to submit their papers and promised to send further details to all participants.

LIST OF ATTENDEES

Mr. James D. Baker
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, Va. 22209

Ms. Annette Basden
Code L51A
Naval Air Station
Pensacola, Florida 32512

Dr. Ralph R. Canter
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, Va. 22209

Dr. Glenn L. Bryan,
ONR Code 450
800 N. Quincy St.
Arlington, Va 22217

Dr. James A. Caplan
U.S. Civil Service Commission
1900 E. St. N.W.
Washington, D.C. 20415

Ms. Cynthia L. Clark
U.S. Civil Service Commission
1900 E St. N.W.
Washington, D.C. 20415

Dr. Norman Cliff
Department of Psychology
University of Southern California
University Park
Los Angeles, California 90007

Dr. Charles H. Cory
Navy Personnel Research and
Development Center
San Diego, CA 92152

Dr. Dorothy D. Edwards
American Institutes of Research
3301 New Mexico Ave, N.W.
Washington, D.C. 20016

Mr. Kenneth I. Epstein
U.S. Army Research
Institute
1300 Wilson Boulevard
Arlington, Va. 22209

Dr. Marshall J. Farr,
ONR Code 458
800 N. Quincy St.
Arlington, Va. 22217

Dr. Victor Fields
13905 Northgate Drive
Silver Spring, Md. 20906

Dr. M. A. Fischl
U.S. Army Research
Institute
1300 Wilson Boulevard
Arlington, Va. 22209

Dr. Dexter Fletcher
U.S. Navy Personnel
Research & Development Center
Code 31
San Diego, Calif. 92152

Dr. Paul Foley
Code 310
Navy Personnel Research
and Development Center
San Diego, Ca. 92152

Mr. Edmund F. Fuchs
2206 Westview Court
Silver Spring, Md. 20910

CMDR. Thomas J. Gallagher
Bureau of Medicine & Surgery
Bldg. 3, Potomac Annex
23rd & E Sts., N.W.
Washington, D.C. 20372

Dr. Robert Gettelfinger
Educational Testing Service
Princeton, N.J. 08540

Dr. Robert Glaser
Learnimg Research & Development Center
University of Pittsburgh
Pittsburgh, Pa. 15213

Dr. William A. Gorham,
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Peggy Goulding
U.S. Civil Service Commission
1900 E. St., N.W.
Washington, D.C. 20415

Dr. Bert F. Green, Jr.
Department of Psychology
The Johns Hopkins University
Baltimore, Md. 21218

Mr. John F. Gugel
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Duncan N. Hansen
Memphis State University
Memphis, Tennessee 38152

CAPT. Dickie Harris, USAF
AFHRL/PES
Lackland Air Force Base,
Texas 78236

Mr. John Hawk
U.S. Department of Labor
Room 8408
601 D St., N.W.
Washington, D.C. 20213

Ms. Julie Hopson
Naval Aerospace Medical
Research Lab
Naval Air Station
Pensacola, Florida 32512

Dr. Carl J. Jensema
Gallaudet College
Kendall Green
Washington, D.C. 20002

Dr. Robert C. Johnson
Gallaudet College
Kendall Green
Washington, D.C. 20002

Ms. Sally Jones
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Stanley Kalisch
Control Data Corp.
4201 Lexington Ave., N.
AHR-207
Arden Hills, Minnesota 55112

Dr. Michael T. Kane
Education Dept.
State University of New York
Stony Brook, N.Y. 11790

Mr. John D. Kraft
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Frederic M. Lord
Educational Testing Service
Princeton, N.J. 08540

Dr. Clifford E. Lunneborg
Bureau of Testing
University of Washington
Seattle, Washinton 98195

Mr. James R. McBride
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

Mr. Richard H. McKillip
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Mr. George MacReady
Department of Measurement and
Statistics
College of Education
University of Maryland
College Park, Md. 20742

Dr. Charles G. Martin
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Ms. Sherryl May
Learning Research and Development
University of Pittsburgh
Pittsburgh, Pa. 15213

Dr. John Mellinger
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington, VA. 22209

Ms. Francia G. Morhardt
U.S. Postal Service
Selection Methods Branch
Room 2300
475 L'Enfant Plaza, West, S.W.
Washington, D.C. 20415

Dr. Mary Anne Nester
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Harry W. O'Neil
ARPA, 745 Architect Bldg.
1400 Wilson Blvd.
Arlington, Va. 22209

Mr. Herbert Ozur
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Mr. Yosef Pavlov
U.S. Civil Service Commission
1900 E st., N.W.
Washington, D.C. 20415

Mr. Steven M. Pine
Psychology Dept.
University of Maryland
College Park, Md. 20740

Mr. Lewis Price
9761 Good Luck Rd. Apt. 9,
Seabrook, Md. 20801

Dr. Martin Rauch
Chief Psychologist
Ministry of Defense
53 Bonn
Post Box 161
Republic of Germany

Dr. Mark D. Reckase
8 Hill Hall
University of Missouri
Columbia, Missouri 65201

Dr. Marty R. Rockway
AFHRL/PES
Lackland Air Force Base, Texas
78236

Dr. Bruce Rogers
Department of Measurement and
Statistics
College of Education
University of Maryland
College Park MD 20742

Richard Rosenblatt
Research Scientist
Hum RRO
300 N. Washington St.
Alexandria Va. 22304

Mr. Robert M. Ross
U.S. Army Research Institute
1300 Wilson Boulevard
Arlington Va. 22209

Mr. Arnold Rubinstein
Naval Materiel Command, Code NMAT-03424
824 Crystal Plaza #6
2221 Jefferson Davis Highway
Arlington, Va. 20360

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

Dr. Frank L. Schmidt
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Donald J. Schwartz
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Robert G. Smith
R&D Plans Division, Code OP 987 P-10
Office, Chief of Naval Operations
5D772 Pentagon
Washington, D.C. 22305

LT. Col. Henry L. Taylor
ODDR&E
3D129 Pentagon
Washington, D.C. 20301

Dr. Vern W. Urry
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Dr. Lonnie D. Valentine
AFHRL/PES
Lackland Air Force Base,
Texas 78236

LCDR C. L. Walker
Central Test Site Fa PTEP
NAUGMS
Dam Neck, Virginia 23461

Major Brian K. Waters
Department of the Air Force
AFHRL Flying Training Div.
Williams AFB Arizona 85224

Dr. David J. Weiss
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

Dr. Harry Wilfong
Armed Forces Vocational
Testing Group
Randolph Air Force Base, Texas 78148

Dr. Hilda Wing
U.S. Civil Service Commission
1900 E St., N.W.
Washington, D.C. 20415

Mr. Victor Wischert
Educational Testing Service
Princeton, N.J. 08540

☆ U.S. GOVERNMENT PRINTING OFFICE : 1976 O—203-473