A COMPARISON OF THE FAIRNESS OF ADAPTIVE AND CONVENTIONAL TESTING STRATEGIES

Steven M. Pine
and
David J. Weiss

RESEARCH REPORT 78-1
AUGUST 1978

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

Prepared under contract No. N00014-76-C-0244, NR150-383
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Research Report 78-1	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): A Comparison of the Fairness of Adaptive and Conventional Testing Strategies	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Steven M. Pine and David J. Weiss	8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0244
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P.E.:61153N PROJ.:RR042-04, T.A.:RR042-04-01, W.U.:NR150-383
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: August 1978
	13. NUMBER OF PAGES: 30
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

test fairness	differential prediction	tailored testing
selection fairness	adaptive testing	individualized testing
bias	computer simulation	item bias
test bias	monte carlo simulation	peaked tests
differential validity	Bayesian testing	uniform tests

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report examines how selection fairness is influenced by the characteristics of a selection instrument in terms of its distribution of item difficulties, level of item discrimination, degree of item bias, and testing strategy. Computer simulation was used in the administration of either a conventional or Bayesian adaptive ability test to a hypothetical target population consisting of a minority and majority subgroup. Fairness was evaluated by three indices which reflect the degree of differential validity,

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

errors in prediction (Cleary's model), and proportion of applicants exceeding a selection cutoff (Thorndike's model). Major findings are (1) when used in conjunction with either the Bayesian or conventional test, differential prediction increased fairness and facilitated the interpretation of the fairness indices; (2) the Bayesian adaptive tests were consistently fairer than the conventional tests for all item pools above the a=.7 discrimination level for tests of more than 30 items; (3) the differential prediction version of the Bayesian adaptive test produced almost perfectly fair performance on all fairness indices at high discrimination levels; and (4) the placement of subgroup prior distribution in the Bayesian adaptive testing procedure can affect test fairness.

CONTENTS

A Comparison of the Fairness of Adaptive and Conventional Testing Strategies

In a previous report, Pine and Weiss (1976) examined how the item characteristics of a conventional test affect its fairness when used in a selection application. That study was concerned with the effects on fairness of (1) the degree of bias in the test items, (2) level of item discrimination, and (3) the distribution of item difficulties. It was found that when fairness was psychometrically operationalized in several ways, these characteristics of a conventional test influenced its fairness. The implication of these results is that the average item discrimination and the distribution of item difficulties, as well as the distributions of test scores for the subgroups being tested, should be considered when evaluating the fairness of a conventional test used as a selection instrument.

Recently, a new class of potential selection tests, referred to as tailored (Lord, 1970) or adaptive (Weiss & Betz, 1973) tests, has emerged. These tests function quite differently from conventional tests and, consequently, may have quite different fairness properties. In adaptive testing each individual is sequentially administered a subset of items from a larger pool of items; each succeeding item administered is contingent upon the testee's responses to the preceding items (Weiss, 1974). Therefore, each individual in a test population will typically receive a unique test, which differs from the tests administered to other people with respect to its average item discrimination and its distribution of item difficulties. Since Pine and Weiss (1976) have already shown that differences among these psychometric properties can affect the fairness of a test, it is appropriate to examine whether, and to what extent, adaptive testing will decrease or increase test fairness.

Based on the current state of knowledge of the psychometric properties of adaptive tests, there are reasons to believe their use can increase test fairness in several ways. Adaptive tests generally produce smaller standard errors of measurement at the extremes of the ability continuum than do conventional tests (e.g., Lord, 1970; Vale, 1975; Vale & Weiss, 1975). Given the relationship between validity and the standard error of ability estimation (Jensema, 1974), this implies that test validity for low-scoring individuals could be expected to increase with adaptive testing. This would be an important result, since members of minority groups tend to obtain low scores on many selection instruments, and evidence of low validity for a minority subgroup can be interpreted as indicating an unfair selection instrument.

Adaptive tests achieve their higher levels of measurement precision (e.g., Lord, 1970; Vale & Weiss, 1975) and test validity (Jensema, 1974; Vale & Weiss, 1975) by tailoring test difficulty to each testee. That is, in an adaptive test the proportion of items answered correctly by each testee should approach a theoretically optimal level (.50 if correct answers by random guessing are not possible). Given an adaptive test item pool with an adequate range of item difficulties, the proportion of items answered correctly would be similar for members of both minority and majority subgroups. Consequently,

test information or precision would be equal. The result may be equal validities for different subgroups and therefore a potential for reduction in unfairness.

A third possible advantage of using adaptive testing would be in extending the principal of differential prediction to single test items. When conventional tests are used, differential prediction involves using separate within-group regression equations to predict the criterion performance of minority and majority subgroups. The logic behind this procedure is that the best prediction of criterion performance for a given subgroup is obtained by developing the prediction equation based only on data from the subgroup for which predictions are to be made. The logical extension of this procedure would be to predict criterion performance using test items which have been calibrated separately for each subgroup. Adaptive testing can accomplish this and, at the same time, adapt the difficulty of the test to the ability level of the examinee.

The purpose of the present study was to compare the properties of one adaptive testing procedure, Owen's (1969) Bayesian adaptive method, with the conventional tests previously studied by Pine and Weiss (1976). Specifically, the investigation was concerned with (1) how item pools with varying item parameters and degrees of item bias would interact with the two test models to affect test fairness, (2) how the use of differential prediction within the context of each testing strategy affected test fairness, and (3) how the placement of the prior ability distribution and choice of a termination criterion affected fairness for the Bayesian strategy.

Method

Assumptions

The above questions were investigated in the context of the same selection situation assumed in the previous study (Pine & Weiss, 1976). The selection process was modeled by a monte carlo simulation; it consisted of administering a selection test to each hypothetical person and using the score from that test to predict an external criterion represented by generated values of the known latent trait, θ. The selection test was assumed to be completely described in terms of its latent trait parameters so that each of its items could be described in terms of its item discrimination (a), item difficulty (b), and probability of being answered correctly by chance guessing (c). Some of the items in the test, however, were assumed to be biased against the minority subgroup; and the degree of item bias was expressed in terms of the latent trait item parameters.

A testee's true ability level on the underlying latent trait, θ, was represented by a number randomly generated from a standard normal distribution. The same θ values were used for both the minority and majority subgroups. The only distinction between the subgroups was how item responses were simulated. For the testees in the "minority" subgroup, the degree of item bias (see Pine & Weiss, 1976, p. 8) was added to the item difficulty to reflect the fact that biased items are effectively more difficult for testees in minority subgroups. This had the effect of lowering the probability of a correct response on biased items for minority subgroup testees.

Tests

Each simulated testee was administered 18 conventional and 9 Bayesian adaptive tests constructed from eighteen 100-item pools, described in Table 1. Table 1 gives the specifications of each pool in terms of its latent trait item parameters a and b (c was assumed to be .20 for all items). Each item in each pool was additionally assumed to have a given level of subgroup bias, which was defined as the difference between the item difficulty (b) parameters for a majority (maj) and minority (min) subgroup. In the case of the conventional tests, items were taken sequentially from all 18 pools; for the Bayesian adaptive tests, items were selected in accordance with Owen's (1969) Bayesian item search algorithm from only the 9 pools having a uniform distribution of difficulties.

Table 1
Distributions of Item Difficulties, Levels of Item Discrimination (a), and Degree of Item Bias in the Simulated Item Pools

Item Pool					
No.	Difficulty Distribution	No.	Difficulty Distribution	a	Bias (b maj-b min)
1	Peaked	10	Uniform	.30	.5
2	Peaked	11	Uniform	.30	1.0
3	Peaked	12	Uniform	.30	2.0
4	Peaked	13	Uniform	.70	.5
5	Peaked	14	Uniform	.70	1.0
6	Peaked	15	Uniform	.70	2.0
7	Peaked	16	Uniform	1.10	.5
8	Peaked	17	Uniform	1.10	1.0
9	Peaked	18	Uniform	1.10	2.0

Adaptive test. The Bayesian adaptive testing strategy (McBride & Weiss, 1976; Owen, 1969; Urry, 1977) begins with an initial (prior) estimate of θ. In this study a normally distributed prior distribution having a mean of 0 and a standard deviation of 1.0 was used. The item to be administered to a testee, described by its latent trait item parameters, is the item that minimizes the expected error function $(\hat{\theta}-\theta)^2$, where $\hat{\theta}$ is the current estimate of ability and a function of the item parameters. Based on the current value of θ, the item parameters, and whether the response to the administered item was correct or incorrect, the current θ estimate is updated. This new estimate then becomes the current estimate, and the cycle is repeated until a termination criterion is reached. In this study Bayesian adaptive testing was terminated when a fixed number of items was administered. Consequently, the standard error of the Bayesian ability estimate varied for testees of different θ levels. The average standard errors of these ability estimates was an additional dependent variable studied for the Bayesian testing strategy; it was compared to the theoretical value based on the fixed number of items administered (Jensema, 1974; Urry, 1977).

Criterion Prediction

Each test was scored in two ways for predicting criterion performance on θ. For the conventional tests, the regression equations for either the majority or minority subgroup were used to convert the total number correct score to the θ metric; for the Bayesian adaptive tests, criterion performance was predicted using Bayesian scoring and either the majority or minority subgroup (i.e., biased) item parameters. When only the majority group regression equations or item parameters were used to estimate θ, this was referred to as the *majority prediction* condition. It was contrasted with the *differential prediction* condition, which used the appropriate subgroups' regression equation or item parameters to estimate θ. In addition, in order to study the effect of test length, each test was scored after 10, 30, and 50 items had been administered.

Fairness

Similar to the previous study (Pine & Weiss, 1976, pp. 9-12), selection fairness was evaluated by R, the correlation between the predicted and true ability on the latent ability; C, the difference between the mean ability levels of the minority and majority subgroups; and T, the difference between the proportion of individuals exceeding the selection cutoff (set at the mean of the majority subgroup) in the two subgroups. In addition, a number of standard distributional statistics were also studied. These included the mean, standard deviation, skewness, and kurtosis of the ability estimates ($\hat{\theta}$) and, for the adaptive testing strategy, the standard error of its ability estimates.

RESULTS

Distributions of Predicted Scores

Means, standard deviations, skewness, and kurtosis indices of ability estimates as a function of the experimental conditions are given for 50-item tests in Table 2; results for tests with 10 and 30 items, which generally parallel those for 50 items, are given in Appendix Tables A and B. In these tables the statistics for the true ability distribution (θ) are given in the first row of the table, listed under the "True" group heading. Only standard deviations are reported for conventional tests in the differential prediction condition, since the other distributional statistics are not affected by differential prediction. Table 2 also gives the results from the subcondition of Bayesian adaptive testing in which the mean of the assumed prior distribution used for the minority subgroup was varied. These conditions are based upon the a=1.1, bias=1.0 condition; in these cases the mean of the prior distribution was set at $\bar{\theta}$=-1.0, -.25, or +1.0.

Means

As Table 2 shows, increasing item bias caused the mean of the minority subgroup to be underpredicted for all of the majority prediction conditions, regardless of the testing strategy employed. In the majority prediction situation, this underprediction increased both with increasing item bias and with increasing item discriminations. For low item discriminations (a=.30) and for the first two levels of bias (0.5 and 1.0), adaptive testing led to a larger

Table 2

Mean and Standard Deviation of Ability Estimates for 50-Item Uniform (U) and Peaked (P) Conventional Tests and the Bayesian Adaptive Test (BAT) as a Function of Item Discrimination (a) and Degree of Item Bias, Using Majority and Differential Prediction, for Majority (maj) and Minority (min) Subgroups

			Mean				Standard Deviation					
			Majority Prediction			Differential Prediction	Majority Prediction			Differential Prediction		
a	Bias	Group	U	P	BAT	BAT	U	P	BAT	U	P	BAT
		True	-.07	-.07	-.07	-.07	1.01	1.01	1.01	1.01	1.01	1.01
.30	0.0	maj	-.07	-.07	-.25	-.25	.80	.81	.80	.80	.81	.80
	0.5	min	-.39	-.40	-.57	-.25	.82	.81	.77	.81	.82	.79
	1.0	min	-.71	-.74	-.86	-.24	.82	.82	.74	.81	.82	.79
	2.0	min	-1.34	-1.36	-1.43	-.23	.82	.80	.72	.82	.82	.83
.70	0.0	maj	-.07	-.07	-.23	-.23	.94	.95	.93	.94	.95	.93
	0.5	min	-.50	-.54	-.64	-.22	.94	.94	.87	.94	.95	.91
	1.0	min	-.95	-.97	-1.04	-.22	.93	.90	.83	.93	.94	.91
	2.0	min	-1.75	-1.71	-1.76	-.23	.82	.72	.71	.92	.90	.92
1.1	0.0	maj	-.07	-.07	-.19	-.19	.97	.96	.92	.97	.96	.92
	0.5	min	-.54	-.54	-.62	-.20	.98	.94	.86	.97	.96	.93
	1.0	min	-1.02	-.96	-1.02	-.20	.95	.85	.78	.96	.94	.90
	2.0	min	-1.85	-1.57	-1.67	-.20	.81	.54	.60	.94	.84	.90
	Bayesian Priors											
1.1	-1.0	min			-1.20	-.28			.90			.98
	-.25	min			-1.10	-.22			.84			.93
	+1.0	min			-.74	-.02			.66			.78

underprediction than did the conventional tests. When differential prediction was used, the adaptive test resulted in substantially less underprediction than did either of the conventional tests. Furthermore, under differential prediction, the degree of underprediction produced by the adaptive test decreased with increasing item discrimination (a) levels.

The mean of the assumed prior ability distribution influenced the predicted mean ability levels in the adaptive test. This effect was substantially less when differential prediction was used. In the majority prediction situation, a prior of θ=1.0 increased the underprediction and a prior of +1.0 decreased it. The smallest degree of underprediction across all conditions was obtained when the prior was set at θ=+1.0; using differential prediction, underprediction was reduced to nearly zero with this prior ability estimate.

Standard Deviations

The standard deviation of the ability distribution (1.01) was underpredicted both for the majority and minority subgroups in all testing conditions. The effect of item discrimination on the standard deviation is reflected by the values for the majority subgroup, where item bias=0. For the conventional tests, the standard deviations increased as the item discriminations were increased from a=.30 to a=1.1. For the adaptive test, there was no increase for levels of item discrimination beyond a=.70. Within each level of item discrimination beyond a=.70, the standard deviations decreased as the item bias was increased, for all testing conditions. The reduction in the standard deviations which resulted from increased item bias was more pronounced at the highest level of item discrimination. The uniform tests reflected this trend the least, and the peaked tests reflected it the most.

The same general trends with respect to the influence of item discrimination and bias on the standard deviations of the distributions of ability estimates occurred when differential prediction was used. However, the influence of item bias was much less in this condition, particularly under adaptive testing, where the overall size of the standard deviations increased relative to the values obtained in the majority prediction condition. For example, where a=1.1 and item bias increased from .5 to 2.0, the adaptive test had standard deviations of .86, .78, and .60 in the majority prediction condition; with differential prediction, a corresponding increase in bias produced standard deviations of .93, .90, and .90. The placement of the prior distribution in the adaptive test influenced the size of the standard deviation in both the majority and differential prediction conditions. In both cases the +1.0 mean prior produced a smaller standard deviation than did the -1.0 mean prior.

Skewness and Kurtosis

For all testing conditions, the degrees of skewness and kurtosis tended to increase in a positive direction as both item discrimination and item bias increased (see Table 3). Positive values of skewness indicate that the mode of the distribution is lower than its arithmetic mean, while positive values of kurtosis indicate that the distribution is more peaked than a normal distribution. In the majority prediction condition, the adaptive test produced score distributions that were always more positively skewed and peaked than any of the uniform conventional tests. Compared to the peaked tests, however, the adaptive test was more positively skewed and peaked only at the lower levels of item bias. In

the combined high item discrimination and high bias conditions, the peaked tests were considerably more leptokurtic than either the uniform conventional or the adaptive tests.

Table 3
Skewness and Kurtosis of Distribution of Ability Estimates for 50-Item Uniform (U) and Peaked (P) Conventional Tests and the Bayesian Adaptive Test (BAT) as a Function of Item Discrimination (a) and Degree of Item Bias, Using Majority and Differential Prediction, for Majority (maj) and Minority (min) Subgroups

			Skewness				Kurtosis			
			Majority Prediction			Differential Prediction	Majority Prediction			Differential Prediction
a	Item Bias	Group	U	P	BAT	BAT	U	P	BAT	BAT
		True	-.01	-.01	-.01	-.01	.22	.22	.22	.22
.30	0.0	maj	.03	-.11	.30	.30	.00	-.06	.06	.06
	0.5	min	.04	.01	.30	.24	-.09	.00	.04	-.08
	1.0	min	.08	.10	.27	.22	-.06	-.02	.15	.07
	2.0	min	.28	.33	.34	.27	.10	.17	.14	.15
.70	0.0	maj	-.08	-.11	.31	.31	-.27	-.66	.01	.01
	0.5	min	.10	.19	.29	.26	-.33	-.66	-.06	-.05
	1.0	min	.31	.49	.40	.26	-.19	-.38	.10	-.02
	2.0	min	.57	1.13	.70	.39	.08	1.08	.48	.01
1.1	0.0	maj	-.03	-.10	.47	.47	-.25	-1.13	.29	.29
	0.5	min	.20	.36	.51	.32	-.24	-.93	.04	-.11
	1.0	min	.30	.86	.62	.34	-.33	-.07	.05	-.19
	2.0	min	.77	2.13	1.12	.49	.42	5.03	1.66	.07
	Bayesian Priors									
1.1	-1.0	min			.34	.14			-.19	-.21
	-.25	min			.46	.27			-.17	-.24
	+1.0	min			.67	.62			.23	.05

The same general trends with respect to the influence of item discrimination and bias on skewness and kurtosis occurred when the differential prediction version of the adaptive test was used. However, as for all the other distributional statistics, differential prediction greatly reduced the influence of item bias and discrimination on skewness and kurtosis. For example, using differential prediction the skewness and kurtosis of the distribution were always equal to or lower than those obtained in the majority prediction condition. Comparing the influence of the +1.0 prior resulted in a more positively skewed distribution, as well as a more leptokurtic distribution, than did the -1.0 mean prior.

Validity

Majority Prediction Condition

Subgroup validities. The validity coefficients (i.e., the correlations between true and estimated ability levels) for the uniform and peaked conventional tests and the adaptive tests, for all experimental conditions, are shown in

Table 4

Validity Correlations for Uniform (U) and Peaked (P) Conventional Tests and for the Bayesian Adaptive Test (BAT) as a Function of Item Discrimination (a) and Degree of Item Bias Using Majority and Differential Prediction for Majority (maj) and Minority (min) Subgroups, and Differences (Diff) Between Subgroups at Test Lengths of 10, 30, and 50 Items

			Majority Prediction									Differential Prediction		
			10 Items			30 Items			50 Items			10 Items	30 Items	50 Items
a	Item Bias	Group	U	P	BAT	U	P	BAT	U	P	BAT	BAT	BAT	BAT
.30	0.0	maj	.493	.540	.501	.725	.741	.709	.793	.802	.772	.501	.709	.772
	0.5	min	.492	.543	.526	.741	.754	.724	.800	.814	.774	.494	.701	.776
		Diff	-.001	.003	.025	.016	.013	.014	.008	.013	.002	-.007	-.009	.005
	1.0	min	.512	.554	.548	.743	.763	.724	.805	.817	.785	.400	.676	.790
		Diff	.019	.014	.047	.018	.022	.014	.012	.016	.013	-.051	-.033	.019
	2.0	min	.523	.540	.570	.749	.759	.733	.810	.811	.797	.543	.721	.803
		Diff	.030	-.001	.069	.024	.019	.024	.017	.009	.026	.042	.011	.032
.70	0.0	maj	.745	.763	.797	.899	.912	.907	.935	.941	.939	.797	.907	.939
	0.5	min	.744	.797	.791	.898	.918	.909	.934	.943	.940	.786	.905	.940
		Diff	-.001	.014	-.006	-.000	.006	.001	-.001	.002	.000	-.010	-.003	.001
	1.0	min	.764	.801	.806	.891	.918	.913	.928	.936	.940	.804	.917	.943
		Diff	.019	.018	.009	-.007	.006	.005	-.006	-.005	.000	.016	.003	.005
	2.0	min	.773	.756	.788	.880	.861	.903	.915	.891	.928	.812	.911	.935
		Diff	.027	-.026	-.009	-.014	-.051	-.005	-.020	-.050	-.011	.016	.003	-.005
1.1	0.0	maj	.820	.869	.881	.932	.940	.956	.961	.954	.968	.881	.956	.968
	0.5	min	.829	.880	.865	.937	.941	.947	.959	.951	.961	.886	.951	.967
		Diff	.009	.011	-.016	.004	.001	-.009	-.002	-.002	-.006	.005	-.005	-.001
	1.0	min	.844	.853	.869	.932	.921	.947	.954	.931	.958	.880	.956	.966
		Diff	.024	-.016	-.012	-.001	-.019	-.009	-.007	-.022	-.010	-.001	.000	-.002
	2.0	min	.824	.753	.840	.915	.818	.929	.934	.831	.933	.877	.946	.955
		Diff	.004	-.115	-.042	-.017	-.122	-.027	-.028	-.123	-.034	-.004	-.010	-.013
	Bayesian Priors													
1.1	-1.0	min			.880			.956			.966	.862	.958	.969
		Diff			-.001			.000			-.002	-.019	.002	.002
	- .25	min			.884			.954			.963	.874	.956	.965
		Diff			.003			-.002			-.005	-.007	-.000	-.002
	+1.0	min			.857			.939			.956	.869	.947	.958
		Diff			-.024			-.017			-.012	-.012	-.009	-.010

Table 4. The three rows in Table 4 labeled "maj" give the validities for the majority subgroup for each value of item discrimination (a). These results correspond to the case in which item bias is zero. The rows labeled "min" and "Diff" give, respectively, the validities for the minority subgroup and the difference between corresponding majority and minority values, for each combination of item discrimination and item bias. In the half of the table labeled "Differential Prediction," only the validity values from the adaptive tests are given. This is because differential prediction in the conventional testing condition amounts to a linear transformation of the test scores and therefore would not change the correlation coefficients. The last six rows of the table give the results from the subcondition of the adaptive test in which the mean of the assumed prior used for the minority subgroup was varied.

The validities for adaptive tests increased with increasing test length and item discrimination. For instance, the lowest validity, r=.501, occurred for a 10-item test with a=.30 and the highest, r=.968, was for a 50-item test with a=1.1. A comparison of corresponding validities between the adaptive test and either type of conventional test for 10-item tests showed that the validities of the adaptive tests were higher in almost all cases in which item discrimination was .70 or higher. For example, for a 10-item test with a=1.1 and item bias of 0.0, the validity correlations were .869 for the peaked test, .820 for the uniform test, and .881 for the adaptive test. For 30- and 50-item tests, the adaptive test had a lower validity for many of the lower discriminating items; but at a=1.1 the adaptive test produced consistently higher validities for all item pools.

Differential validity. A major concern with respect to test fairness is not just how validity varies as a function of the test characteristics for a given subgroup but, more importantly, how validity varies differentially between subgroups. The reason for this is that if a difference in subgroup validities does exist, the predictions made on the basis of the test scores are not as accurate for one subgroup as for the other. Therefore, the effect of item bias on validity was studied by comparing the validities for both subgroups for all the item pools and test lengths. To facilitate this analysis, differences between subgroup validities were determined. Differential validity was thus defined as

$$r_{diff} = r_{min} - r_{maj}. \qquad [1]$$

A negative value of differential validity indicates that the majority subgroup had a higher validity coefficient than the minority subgroup. These values appear in Table 4 in the rows designated "Diff."

Table 4 shows that for the lowest a value (a=.3) as item bias increased, validity differences increased for the uniform and adaptive tests but decreased for the peaked test. Also, at a=.30 differential validity tended to be positive (i.e., minority subgroup validities were higher for all test types), with the largest values tending to occur for the adaptive test. However, for item discriminations of a=.70 and 1.1 for test lengths of 30 and 50, the direction of differential validity reversed, so that higher validity correlations were observed for the majority subgroup. As the degree of item bias and item discrimination increased, the size of the negative difference became substantial for the peaked test relative to the adaptive test, while the uniform tests

generally produced a slightly smaller negative validity difference than did the adaptive test. For example, the 50-item peaked test with a=1.1 and bias of 2.0 had a -.123 difference between the subgroup validities, compared to values of -.028 for the uniform test and -.034 for the adaptive test.

The effect of choice of prior distribution on the validity of the adaptive procedure was that when priors other than the majority subgroup prior were used, validity tended to increase as the priors became higher in negative value. Since the priors were varied only for the minority subgroup, the effect on differential validity (i.e., the difference between majority and minority subgroup validity coefficients) was that the negative priors produced differential validities closer to zero than in the positive prior case.

Differential Prediction Condition

Subgroup validities. At item discrimination levels beyond a=.70, the differential prediction version of the adaptive test produced higher validities than did the majority version of the adaptive test or either conventional test. This relative advantage increased as item bias increased. Typical are the values for a 50-item adaptive test with a=1.1 and bias of 1.0, where r=.966 under differential prediction and r=.958 under majority prediction; this can be compared with r=.954 and r=.931, respectively, for the uniform and peaked conventional tests. Under these same conditions, the validity for the adaptive test increased to r=.968 by using the -1.0 mean prior for the minority subgroup.

Differential validity. Validity differences between subgroups were reduced by using the differential prediction version of the adaptive test. Unlike the majority prediction case, in which the uniform conventional tests often showed less differential validity than the adaptive test, in this condition the adaptive test generally had the smaller differential validity. Furthermore, the advantage of the adaptive test increased as item bias increased. The negative mean priors tended to increase differential validity for the 10-item test but had relatively little effect on the longer tests. The +1.0 mean prior, however, led to an increase in differential validity at all three test lengths.

C-Fairness

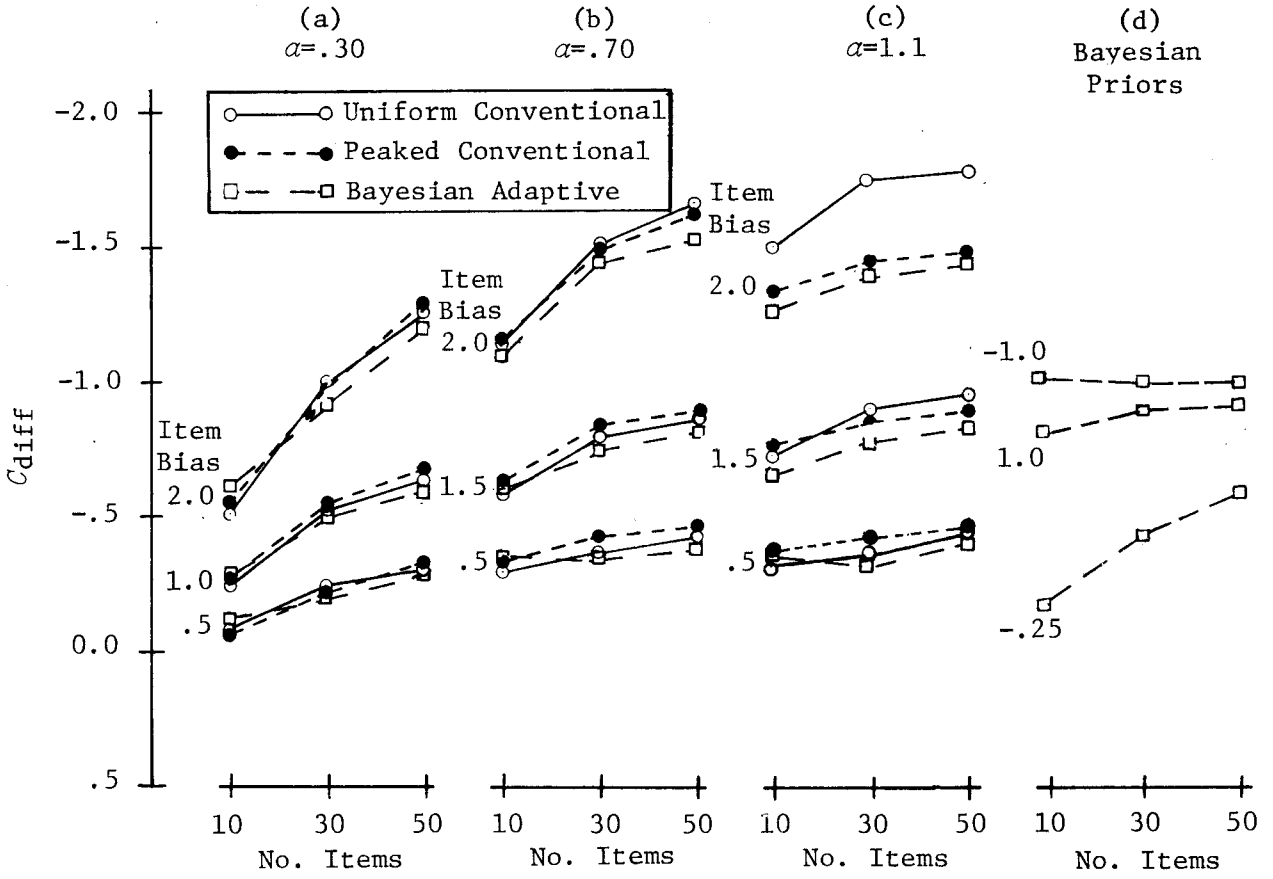
Majority Prediction

The Cleary-type fairness measure, C, was defined as the mean of the predicted ability, $\bar{\hat{\theta}}$, minus the mean of the true ability, $\bar{\theta}$ (Pine & Weiss, 1976, p. 10). Therefore, the C-index is in the same units as θ, which had a mean of 0 and a standard deviation of 1.0.

In Figure 1 values of "C_{diff}," the subgroup differences in the C-indices ($C_{min} - C_{maj}$), are plotted against test length for the uniform, peaked conventional, and adaptive tests for all levels of item discrimination in the majority prediction condition. (Numerical values of C by subgroup are given in Appendix Table C.) It can easily be seen by some simple algebraic manipulation (substituting $[\bar{\hat{\theta}}_{min} - \bar{\theta}_{min}]$ for C_{min}, $[\bar{\hat{\theta}}_{maj} - \bar{\theta}_{maj}]$ for C_{maj}, and subtracting) that

$C_{diff} = \bar{\hat{\theta}}_{min} - \bar{\hat{\theta}}_{maj}$ (recall that both subgroups had the same mean true ability level, $\bar{\theta}$). Therefore, a negative value of C_{diff} implies unfairness to the minority subgroup in the sense that their mean ability is underpredicted relative to the predicted mean ability of the majority subgroup.

Figure 1
Group Differences in C-Index (C_{diff}) as a Function of Item Discrimination (a), Item Bias, and Test Length, Using Majority Prediction



The C-index indicated increased unfairness for the majority subgroup (higher negative values of C_{diff}) as item bias was increased from .5 to 2.0. This trend increased as a negatively accelerating function of test length, with the rate of increase varying as a function of item discrimination and degree of item bias. For all tests, increasing item bias tended to be associated with higher levels of C_{diff} for longer tests within a level of item discrimination. This effect of test length decreased, however, as item discrimination increased.

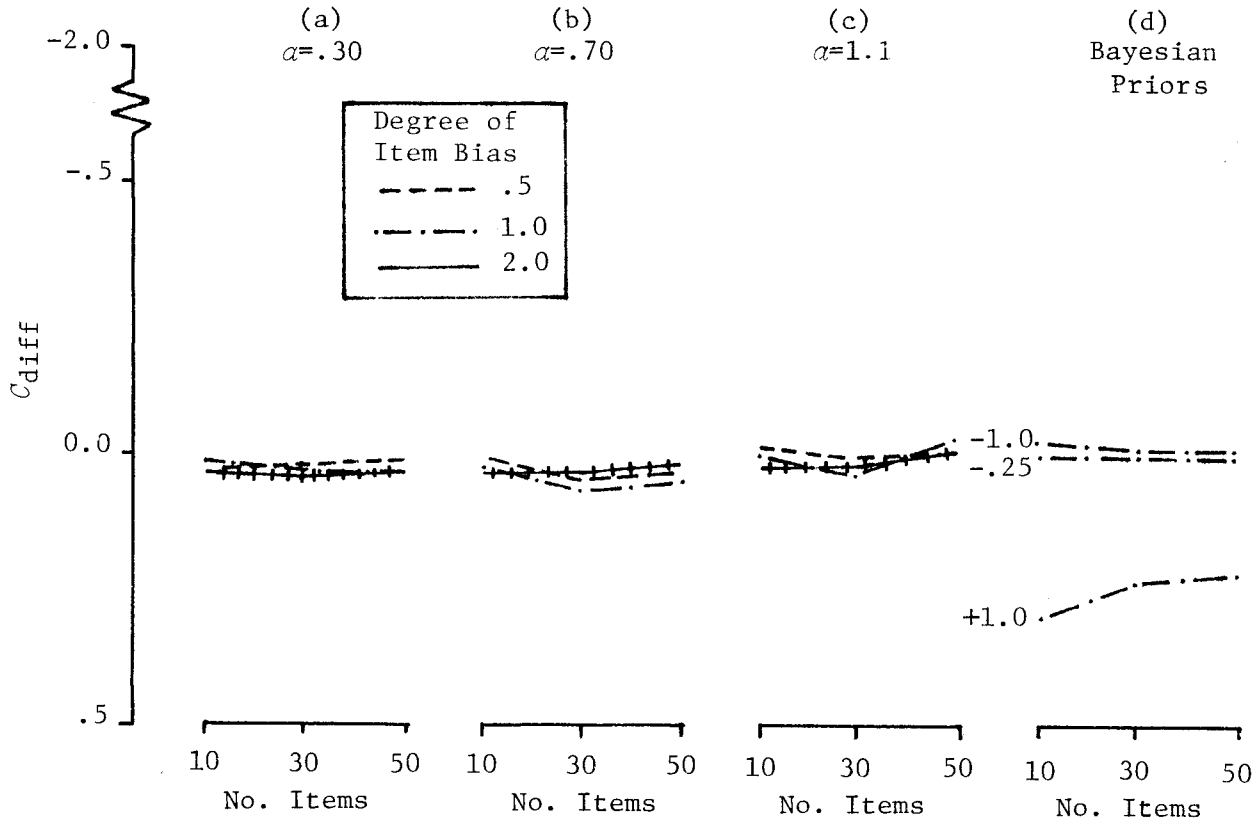
There were small differences between tests on C_{diff} at the a=.30 level of item discrimination. For 30- and 50-item tests, the adaptive test generally had lower levels of C_{diff}. However, at the higher levels of item discrimination

and item bias, the adaptive test showed substantial advantage over either conventional test in producing lower levels of C_{diff}. For example, for a 30-item test with a=1.1 and item bias of 2.0, the adaptive test produced ability estimates one-third of a standard deviation less biased than those produced by the uniform conventional test.

The choice of a prior distribution for the minority subgroups in the adaptive test directly affected the resulting values of C_{diff}. The larger the mean of the prior ability distribution (in a positive direction), the lower the values of C_{diff}. Increasing test length had the effect of reducing differences in C_{diff} due to the different Bayesian prior ability estimates.

Figure 2

Group Differences in C-Index (C_{diff}) as a Function of Item Discrimination (a), Item Bias, and Test Length, Using Differential Prediction with the Bayesian Adaptive Test



Differential Prediction

Values of C_{diff} are plotted in Figure 2 against test length for the adaptive test for all levels of item discrimination in the differential prediction condition. (Only the results of the adaptive tests are plotted,

since by definition C_{diff} is always equal to zero under differential prediction for conventional tests; see Pine & Weiss, 1976, pp. 10-11.) As Figure 2 indicates, when differential prediction was employed with the adaptive test, differences in the degree of unfairness between subgroups were practically eliminated, especially at the high levels of item discrimination. For a=.30 there was a tendency for the minority subgroup to be overpredicted (i.e., positive values of C_{diff}). This tendency, however, decreased as item discrimination increased (Figures 2b and 2c). At a=1.1 all C_{diff} values were practically zero, the largest (which occurred at the shortest test length) being -.022.

The relationship of C_{diff} to test length when various priors were used was very similar to that found in the majority prediction case, except that values of C_{diff} were closer together and shifted down to near the zero bias level. Because of this shift, the use of the +1.0 mean prior caused an overprediction of the minority subgroup. Once again, as test length was increased, values of C_{diff} resulting from the differential priors were more similar.

T-Fairness

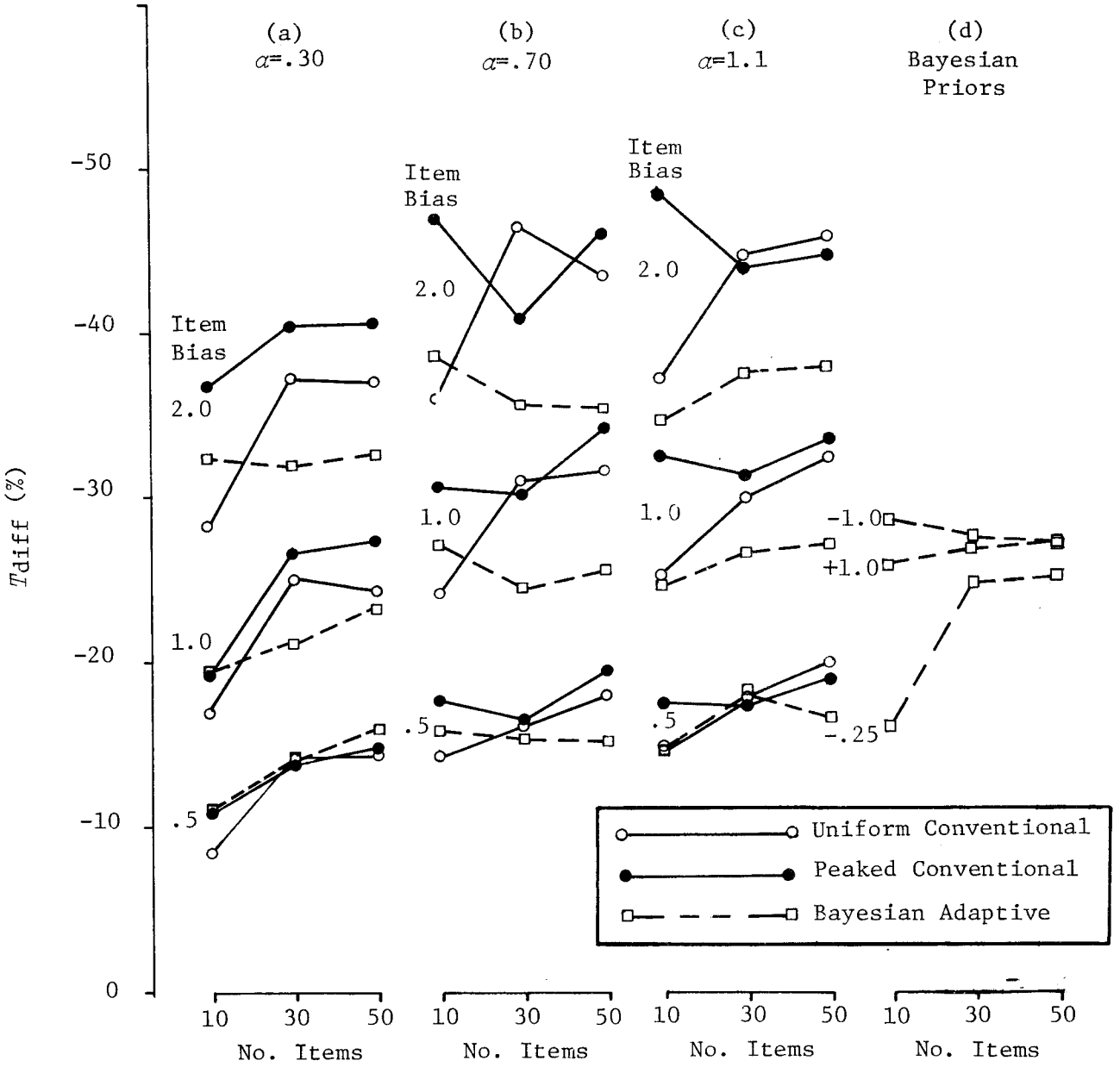
Majority Prediction

T_{diff} can be defined as the differences between the T-indices for the majority and minority subgroups and is equivalent to the difference between the percentage of minority and majority testees predicted to be above the majority group average (see Pine & Weiss, 1976, p. 11). A negative value of T_{diff} indicates that the percent predicted to be above average was smaller for the minority than for the majority subgroup, i.e., the test was more unfair to the minority subgroup. Values of T_{diff} for the conventional and adaptive tests are shown in Figure 3. The numerical values of T by subgroup for all tests are shown in Appendix Table D.

As Figure 3 shows, T_{diff} varied in a complex way as a function of item discrimination, degree of item bias, and test length. The adaptive test showed smaller effects due to increases in test length. Comparing the tests under different degrees of item bias and levels of item discrimination, the 10-item uniform tests were usually fairest (i.e., had smallest values of T_{diff}) when a=.30 and .70, regardless of level of item bias. In all other cases for a=.70 and 1.1, the adaptive test produced levels of T_{diff} closest to zero. At the highest level of discrimination (a=1.1) and bias (2.0) for a 50-item test, T_{diff}=37.6% for the adaptive test, 45.5% for the uniform conventional test, and 44.2% for the peaked conventional test. In terms of the percentage of examinees who would be judged above average, this implies a difference between the adaptive and uniform conventional tests of 7.8% in the number of minority, compared to majority, examinees.

The effect of choosing a negative mean prior in the Bayesian adaptive test was to produce a negative bias against the minority subgroup. Holding test

Figure 3
Group Differences in T-Index (T_{diff}) as a Function of
Item Discrimination (a), Item Bias, and Test Length, Using Majority Prediction



length constant and comparing T_{diff} between levels of prior ability estimates, the differences were about 2% to 3% for 30- and 50-item tests and 9% to 12% for 10-item tests. However, these differences, due to choice of priors, were relatively small compared to the effect of varying item bias in the cases in which a prior equal to the true mean ability was used.

Differential Prediction

The results of using differential prediction in conjunction with the adaptive test on T-fairness are shown in Figure 4. (For simplicity, the results of differential prediction on conventional tests are not shown in Figure 4; but

their numerical values are given, along with those from the adaptive testing condition in Appendix Table D. A detailed discussion of the effects of differential prediction on conventional tests can be found in Pine & Weiss, 1976.) Appendix Table D shows that the main effect of using differential prediction with any of the three testing strategies was that a much larger percentage of minority applicants was predicted to be above average than in the majority prediction condition. Consequently, the general level of unfairness was reduced using differential prediction.

Figure 4
Group Differences in T-Index (T_{diff}) as a Function of Item Discrimination (a), Item Bias, and Test Length, Using Differential Prediction with the Bayesian Adaptive Test

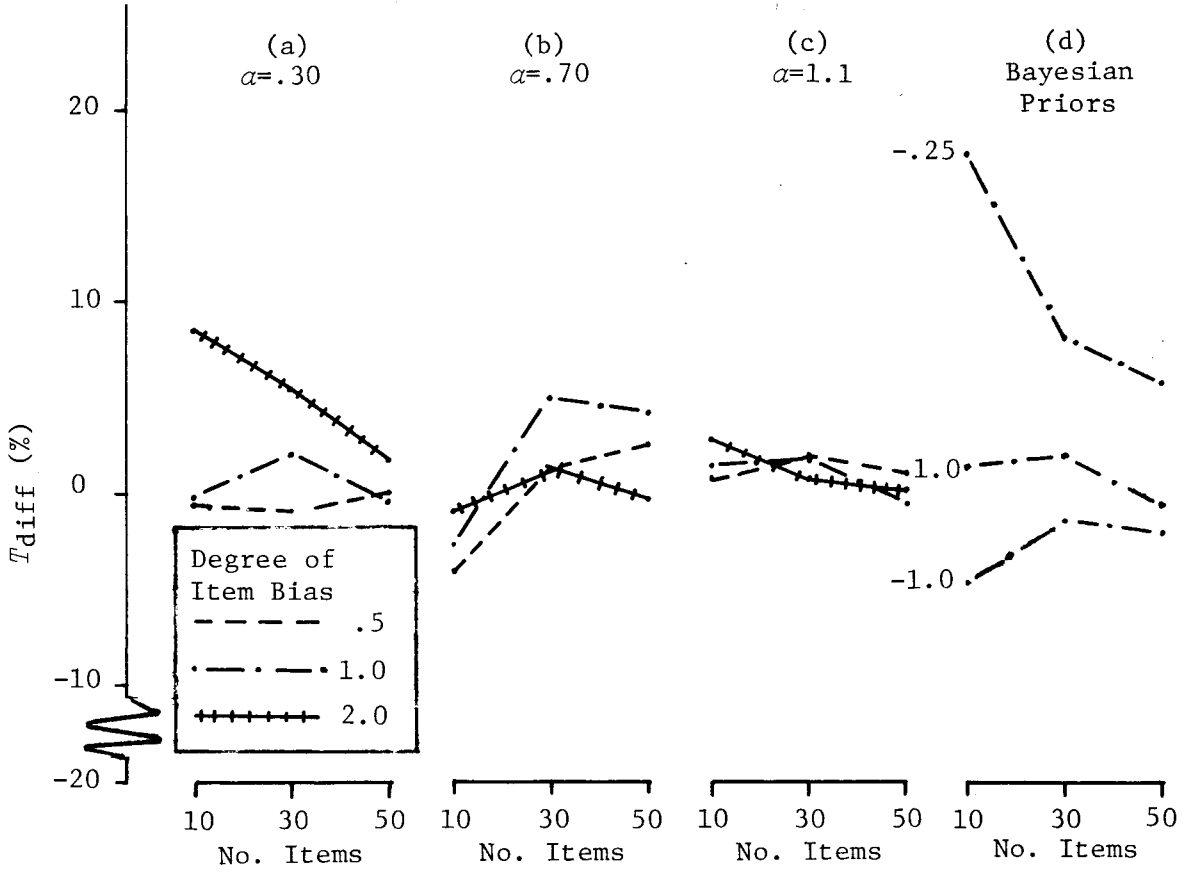


Figure 4 shows that with the differential prediction version of the adaptive test, the minority subgroup tended to have a greater percentage of examinees above the mean than did the majority subgroup. For example, at the a=.30 level of discrimination, a 10-item test having an average item bias of 2.0 had 8.2% more of the minority subgroup above average than the majority subgroup. However, as item discrimination increased, this overprediction was reduced to the point where at a=1.1, T_{diff} values ranged between +2.6% to -.4%.

This overprediction of the minority subgroup never occurred in the majority prediction case for any other of the test strategies (see Figure 3). However,

a similar result did occur when differential prediction was used in conjunction with the uniform conventional tests, but only at the shorter test lengths or at the lowest level of item discrimination. Overprediction of the minority subgroup almost never occurred with the peaked conventional test.

When differential prediction was used, T_{diff} indicated that the adaptive test produced a fairer test than the conventional test in all but 6 cases (out of 18 possible) when discriminations were .30 and .70 (see Appendix Table D). At the high discrimination level (a=1.1), there was only one instance in which one of the conventional tests produced levels of T_{diff} nearer zero than the adaptive test (i.e., a 30-item uniform test with bias of 1.0); and the difference there was small, 1.6% for the uniform test compared to 1.8% for the adaptive test.

Contrary to what was found when majority prediction was used, the choice of priors in the adaptive test had a relatively large effect on T_{diff} compared to item bias. As was the case with the C-index, differential prediction sometimes resulted in a positive bias in favor of the minority subgroup; this occurred primarily when the +1.0 prior was used.

The Standard Error of Estimation in Bayesian Adaptive Testing

The Bayesian adaptive testing procedure provides an estimate of the mean ($\hat{\theta}$) and variance (s_m^2) of the estimated ability distribution after each test item is administered. In the present study, s_m^2 (and its square root, s_m) varied across testees because a fixed test length termination criterion was employed. The average s_m^2 (the standard error of estimate) was computed for each experimental condition and compared to its actual value; the resulting ratios are plotted for majority and differential prediction conditions in Figures 5 and 6, respectively.

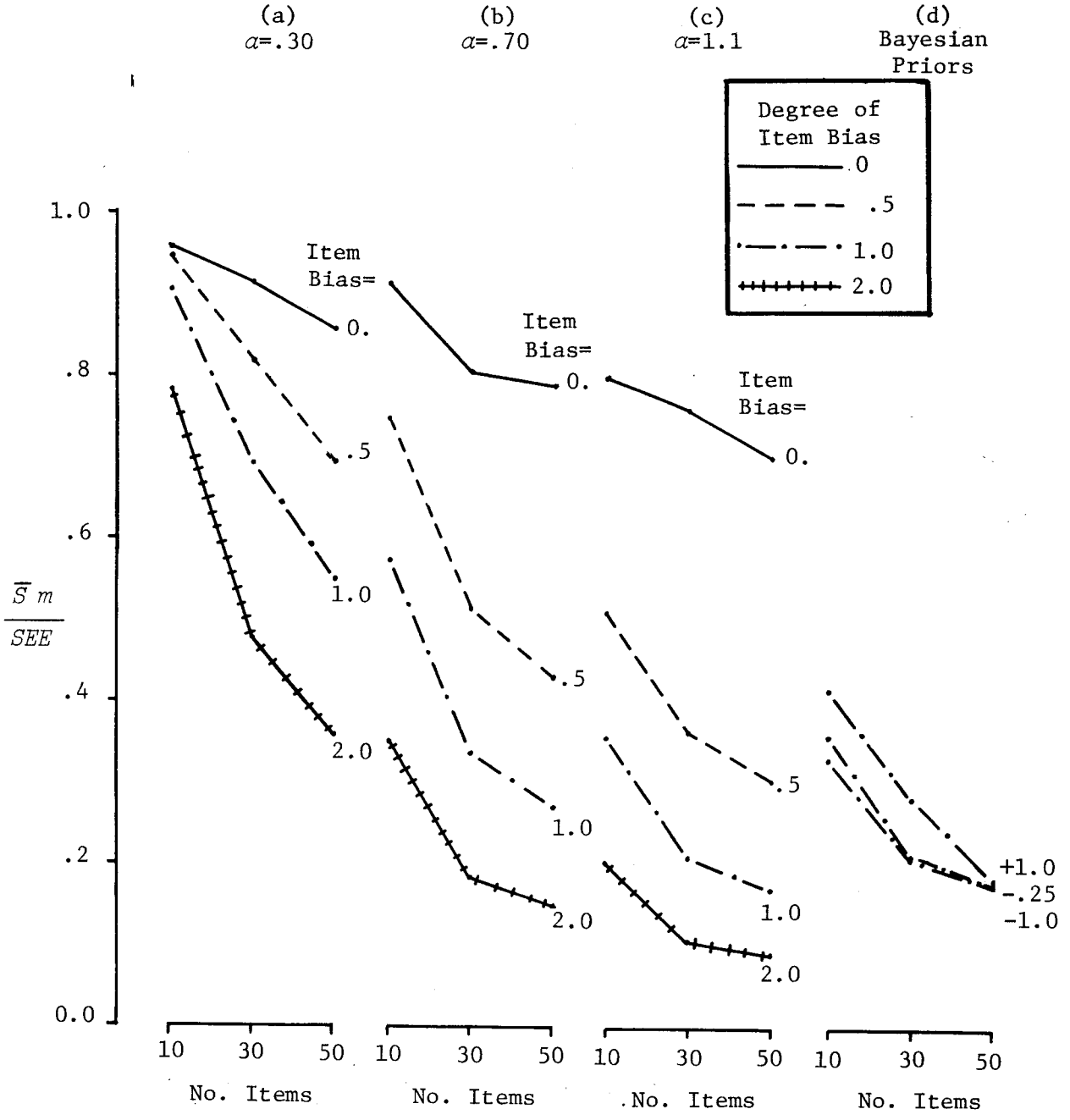
Majority Prediction

The average posterior standard deviation ($\bar{s}_m$) after m test items were administered underestimated the theoretical value of the standard error of estimate (*SEE*) in all conditions (see Figure 5). This underestimation became progressively worse as item bias increased. For example, at a=1.1 for a 50-item test with bias of 2.0, the obtained ratio was .094. Little difference in the ratios resulted from use of the negative mean priors. The +1.00 prior, however, produced a larger ratio at each test length.

Differential Prediction

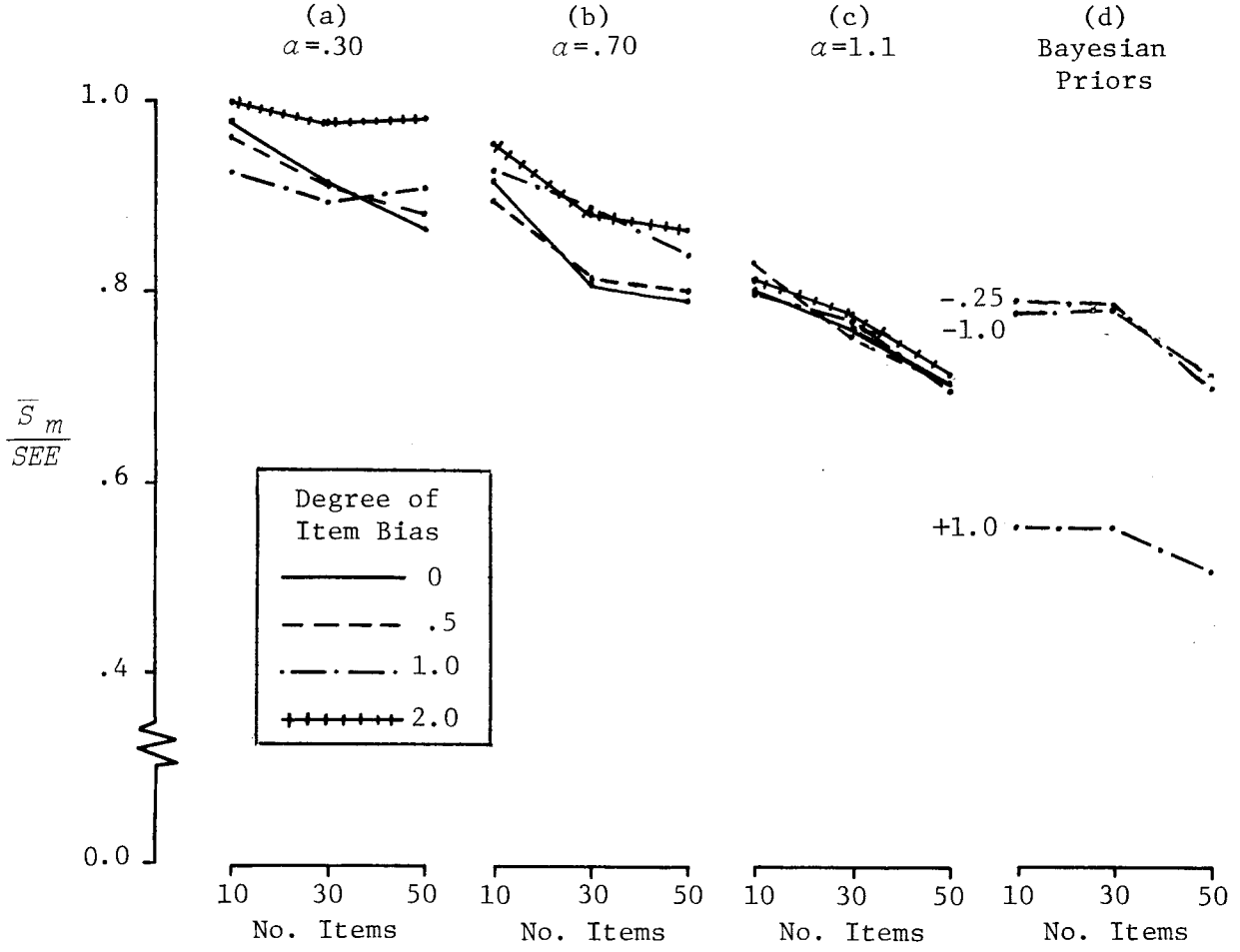
The ratios for the differential prediction case are plotted in Figure 6. In this condition, the effect of item bias on the size of the standard error ratios was greatly reduced. However, a small effect due to bias was still in evidence, particularly at the lower discrimination levels. The effect due to test length was also reduced, particularly at a=.30 (Figure 6a). There was still a systematic decrease in the ratios as a function of increasing discrimi-

Figure 5
Ratio of Estimated Standard Error of Estimate Based on Bayesian Posterior Variance to Predicted Standard Error of Estimate as a Function of Item Discrimination (α), Item Bias, and Test Length, Using Majority Prediction



nation. As in the majority prediction case, there was little difference due to use of negative mean prior ability estimates. The effect of using the +1.0 prior, however, was a relatively larger decrease in the standard error ratio than occurred in the majority prediction case.

Figure 6
Ratio of Estimated Standard Error of Estimate Based on Bayesian Posterior Variance to Predicted Standard Error of Estimate as a Function of Item Discrimination (a), Item Bias, and Test Length, Using Differential Prediction



DISCUSSION

In a previous study (Pine & Weiss, 1976), it was shown that the fairness of a test when used as a selection instrument will depend on the item characteristics of the test items and on the way fairness is defined. The purpose of the present study was to extend these results by investigating the effects on fairness of varying the testing strategy as well as the characteristics of the test items.

Shape of the Predicted Ability Distributions

The shape of the test score distributions varied systematically as a function of the independent variables manipulated in this study. The effects of two of these--level of item discrimination and distribution of item difficulties--have been studied rather extensively in previous research (Cronbach & Warrington, 1952; Lord & Novick, 1968; Pine & Weiss, 1976; Urry, 1969). However, the influence of type of testing strategy (i.e., conventional versus

adaptive) and use of differential prediction on the shape of score distributions has not received previous attention.

Skewness and kurtosis. In general, the findings were that as the degree of item bias and item discrimination increased, the shape of the score distributions became increasingly positively skewed and flat relative to a normal distribution. These findings are consistent with Lord & Novick's (1968, chap. 16) graphical demonstration that increasing test discrimination will tend to flatten the true score distribution, while increasing the difficulty of a test will cause positive skewness in the true score distribution.

The shape of the ability score distributions was also a function of the testing strategy. The peaked conventional tests were more strongly influenced by the presence of item bias and by increasing item discrimination than were the other test types. At the highest degree of bias and discrimination, the score distributions for the peaked conventional tests became exceedingly flat and positively skewed. In contrast, the differential prediction version of the Bayesian adaptive test produced ability distributions relatively unchanged in shape across bias and discrimination conditons.

Means and standard deviations. Both the means and standard deviations of the distributions tended to be underestimated. The underprediction of the standard deviations is the direct result of regression towards the mean. The effect of item bias on the mean of the score distributions was also in the predicted direction, since a direct inverse relationship would be expected between degree of bias in test items and test scores. In terms of the means and standard deviations, the uniform conventional tests showed the least underprediction of the standard deviations, and the peaked conventional tests generally resulted in the smallest underprediction of the means. However, the condition in which values of the prior ability estimates of the adaptive testing strategy were varied produced less underprediction of both the mean and standard deviation of the ability distribution. Furthermore, when the differential version of the adaptive test was employed, even lower levels of underprediction resulted.

Validity Index

The results of the validity data have several implications for the construction of tests and the interpretation of existing test data. First, as was previously discussed (Pine & Weiss, 1976), the validity results offer a possible explanation for the often-reported but controversial phenomenon of differential validity. According to the model used in this study, the existence of subgroup validity differences (i.e., differential validity) is interpreted as an indication of the fairness of a selection instrument. The smaller the difference between validity coefficients, the fairer the selection instrument.

Several researchers (i.e., Campbell, Crooks, Mahoney, & Rock, 1973; Schmidt, Berner, & Hunter, 1973) have presented arguments, based on various analyses of empirical data, that differential validity does not exist as a substantive phenomenon. The results of this study indicate that differential validity is real and, in fact, can be expected when test items are biased against one of the subgroups being tested. Furthermore, based on the present study, it can be seen that the properties of the testing strategy will also

influence differential validity. Both conventional and adaptive testing strategies had a direct influence on the extent to which a given degree of item bias affected differential validity.

Variations within each of these testing strategies also affected differential validity. Within the conventional tests, the distribution of item difficulties (peaked or uniform) and item discrimination level had a differential effect. For the adaptive tests, the influence of item bias on differential validity varied as a function of the level of item discrimination and choice of prior ability estimate, whether majority item parameters or subgroup parameters (i.e., the majority or differential prediction condition) were used.

Conventional versus adaptive tests. Both the majority prediction version of the adaptive test and the uniform conventional tests tended to produce a smaller differential validity for a given degree of item bias than did the peaked conventional test at the longer test lengths and higher levels of item discrimination. The adaptive test and uniform conventional test produced very similar levels of differential validity when majority prediction and a zero prior ability estimate were used. However, the adaptive test produced higher minority subgroup validities and therefore a smaller differential validity when the prior ability estimate used for the minority subgroup was one standard deviation below the mean of the majority subgroup.

The reason that using a negative prior led to higher minority subgroup validity was that in this condition the test items were biased by one standard deviation on the difficulty scale. This resulted in the minority testees responding as though they were one standard deviation below their true ability level. Therefore, the -1.00 prior ability estimate more closely matched their effective mean ability level.

Item discrimination. The influence of item discrimination on differential validity also varied as a function of the testing strategy. In the majority prediction condition, the level of item discrimination which led to the smallest degree of differential validity appeared to be lower than might have been suspected for both conventional and adaptive tests. For the peaked conventional test, the lowest discrimination value led to the least differential validity. However, with both the majority prediction version of the adaptive test and the uniform conventional tests, the intermediate level of item discrimination led to the least differential validity.

That differential validity was not directly related to level of item discrimination may seem surprising, since both in this study and in Urry's (1969) study, it was shown that validity increased with increased item discrimination for the difficulty levels examined. The essential factor, however, is that when validity is considered with regard to fairness (i.e., differential validity), the effect of item bias must be considered. The presence of item bias effectively increases the difficulty parameter (b) for the minority subgroup while leaving the b level unchanged for the majority testees. Since validity is a function of both b and item discrimination (a), the reported effect in differential validity resulted.

It appears that when the same prediction parameters are used in a conventional test for all subgroups, the usual practice of selecting items having the highest discriminations will generally have the effect of increasing subgroup

validity differences if test items are biased. The more biased the items are, the larger will be the difference in subgroup validities. A reduction in validity differences can be achieved by using intermediate levels of item discrimination; of course, then the reduction in differential validity will have been achieved at the cost of lowering the overall level of validity. In some situations, particularly if the majority subgroup validity is relatively high, such a tradeoff may be desirable.

Differential prediction in the adaptive test. The differential prediction version of the adaptive test, however, provides a means of controlling the level of differential validity while maintaining a high level of validity. When this strategy was used, it produced both the smallest differential validity and the highest overall validity for both subgroups. Apparently, in terms of validity, fairness is most readily attainable by adopting the testing strategy which has the greatest ability to adapt to the individual testee. Among the testing strategies investigated in this study, this was the differential prediction version of the Bayesian adaptive test, in which test items were selected for a given testee on the basis of the item parameters derived for the testee's subgroup.

Other Indices of Selection Fairness

In the context of this study, the *C*-index, based on Cleary's fairness model, gave the degree of statistical bias in the estimation of a known value of ability (θ). The *T*-index, based on Thorndike's definition of fairness, reflected the meaning of these mis-estimations in terms of the percentage of applicants who were predicted to exceed some qualifying point of ability--in this case, the mean of the majority population.

The Cleary view of fairness tends to optimize selection from the vantage point of the selecting institution, since it assures that the ablest candidates will be selected. The Thorndike model tends to be more liberal from the viewpoint of the minority subgroup. Even in situations in which the Cleary index indicates a perfectly fair test, it has been previously shown by Schmidt and Hunter (1974) that the Thorndike index may still indicate unfairness. This result was replicated both in the previous study (Pine & Weiss, 1976) and in the present study.

From the previous study it was shown that even within conventional tests, the spread of item difficulties can have a strong effect on fairness at some levels of item discrimination and for some test lengths. For the levels of discrimination and test lengths most commonly found in practice, the general finding was that the peaked test was fairer in terms of the *C*-index and the uniform test was fairer in terms of the *T*-index, when majority prediction was employed. The differential prediction condition indicated the conservative nature of the *C*-index. By definition, in this condition all tests were perfectly fair by the Cleary model; yet the *T*-index indicated the presence of substantial unfairness, particularly for very short tests or for tests composed of highly discriminating items. Furthermore, with differential prediction of ability, the use of tests with uniform distributions of item difficulties was consistently more favorable to the minority subgroup.

Conventional versus adaptive tests. In the present study Bayesian adaptive tests were compared to the same conventional tests used in the previous

study (Pine & Weiss, 1976) in order to determine the effects of testing strategy on test fairness. For the levels of item discrimination and test lengths most commonly found in practice, the general finding was that the adaptive tests were fairer than either the peaked or uniform conventional tests in terms of both the *C*- and *T*-indices when majority prediction was employed. The advantage of the adaptive tests over conventional tests was increased further by adjusting the prior ability estimate used for the minority subgroup.

In the differential prediction condition using the *C*-index, all conventional tests are perfectly fair by definition; therefore, they cannot be improved by adaptive testing. Yet, at high levels of item discrimination, the adaptive test approached this ideal level of performance. On the *T*-index, there was an even greater advantage in favor of the minority subgroup using adaptive tests, as compared to conventional tests, when differential prediction was used rather than majority prediction.

Item discrimination. All testing strategies in the majority prediction condition showed an overall improvement in fairness on the *T*-index as item discrimination decreased. The *C*-index also displayed this relationship and proved to be less affected by increases in item bias at low levels of item discrimination. These findings are disturbing, since it does not seem logical that "poor" items should have to be used in order to achieve fairness, particularly since their use reduces overall validity.

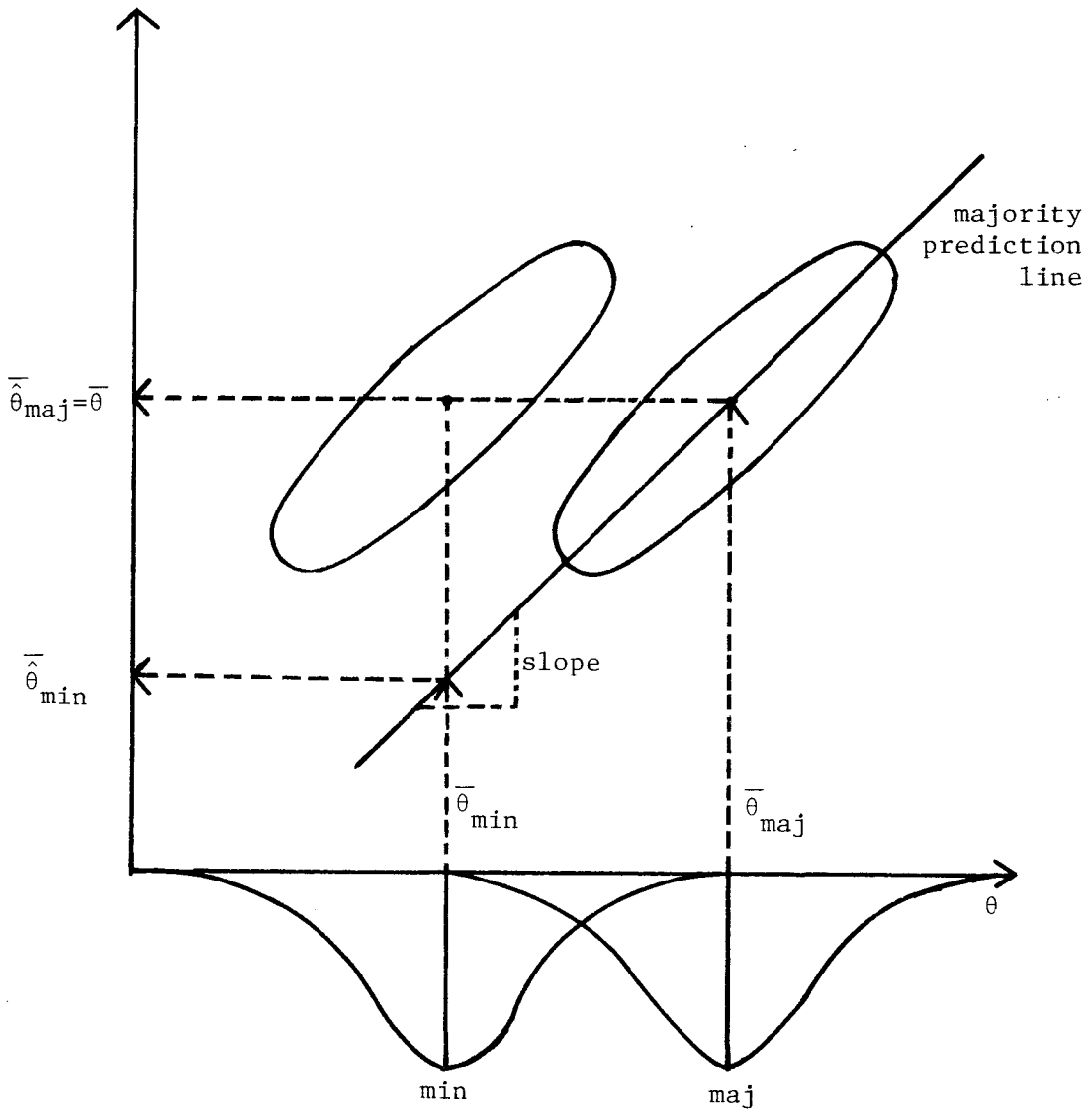
However, this result is an artifact of using the majority prediction parameters for both subgroups with either conventional or adaptive testing strategies. This can be seen in Figure 7. It is obvious from Figure 7 that since the minority subgroup mean is predicted through the majority prediction line, the mean predicted ability level of the minority subgroup will be highest when the slope of the regression line (and therefore *r*) is lowest. Since the levels of item discrimination examined in this report were directly proportional to *r* (the correlation of observed and estimated ability levels), it follows that the predicted mean ability levels of the minority subgroup will increase as item discrimination decreases. Both the *C*- and *T*-indices indicated increased fairness as the mean of the minority subgroup increased. Therefore, it follows that fairness as measured by both the *C*- and *T*-indices should improve as item discrimination is decreased.

Advantages of differential prediction. Another problem which occurs with the use of majority prediction is that the test items which are optimal with respect to reducing differential validity are not the same items that are optimal with respect to the *C*- and *T*-indices of fairness. This dilemma can be resolved with the adaptive testing model by using differential prediction. Overall fairness was optimized in this case in a logical, consistent manner for all three of the fairness indices. With differential prediction, fairness as measured by the *R*-, *C*-, and *T*-indices was dramatically increased for all levels of item discrimination and item bias. Furthermore, adaptive testing displayed decreased sensitivity to increasing item bias with respect to each of the fairness indices as item discrimination was increased.

The effect of differential prediction within the context of the conventional test model could be observed only on the *T*-index, since the *R*-index was unchanged and $C=0$, by definition. Both the uniform and peaked tests showed a marked improvement in fairness under differential prediction, although not as

much as was found with the adaptive test. However, even when differential prediction was employed, the peaked test still was most robust to increasing item bias and fairness overall at the lowest level of item discrimination. Consequently, it would appear that when using conventional tests, a high degree of fairness can be obtained (both with respect to the T-index and differential validity) by peaking item difficulties and using items with a relatively low level of discrimination. This policy, however, would decrease the overall level of test validity.

Figure 7
Linear Prediction of Minority Mean
Ability Using Majority Prediction Line



The uniform conventional tests produced results which were much more similar to those found with the adaptive tests. Differential prediction improved fairness at higher levels of discrimination. This similarity between the uniform and adaptive tests may have resulted from both types of tests having a uniform spread of item difficulties.

Prior ability distributions. The prior ability distribution chosen to begin the Bayesian ability estimation process also appeared to affect the fairness of each test. Prior ability estimates which underestimated the true ability caused an increased underprediction of the mean, compounding the degree of underprediction caused by item bias. This underprediction was directly reflected in the *C*- and *T*-fairness indices.

The choice of a prior ability estimate did not influence differential validity in the same way as it did the *C* and *T* measures. For the R_{diff} index (assuming the presence of item bias), prior ability estimates below the true population ability levels actually improved fairness (i.e., reduced differential validity). This was a result of the fact that the presence of item bias made it appear as though the minority subgroup's mean ability level was lower than it really was; consequently, the low prior ability estimate was effectively more appropriate, producing a higher validity for that subgroup.

The conflicting results of the effect of prior ability estimates on test fairness were not alleviated by using differential prediction. In the differential prediction condition, prior ability estimates higher than true ability levels led to estimation bias, this time in favor of the minority subgroup. The higher priors still, however, led to a less favorable result with respect to differential validity.

There is a need for more research on the best procedures to follow in choosing subgroup prior ability estimates in Bayesian adaptive testing. The problem is that if different priors are used for each subgroup, based on available ability data, certain minority subgroups may be unfairly affected. This might occur for subgroups that have tended to score lower on past tests--which may have been biased. Lower mean prior ability estimates would be used for members of these subgroups which would then lead to lower levels of estimated ability with the Bayesian adaptive testing method. On the other hand, if identical prior ability estimates are used for all subgroups and there is a true ability level difference between the subgroups, those subgroups that are overestimated will be given an unfair advantage. Furthermore, this advantage will be much larger than the disadvantage in fairness that results from using a prior ability estimate which underpredicts true ability. However, data from the current study suggest that both of these undesirable outcomes can be minimized by increasing test length.

The Bayesian Error of Estimate

In the present study the adaptive testing termination criterion was always based on test length. However, the observed posterior standard deviation (s_m), which can be thought of as an estimate of the standard error of the final ability estimate (*SEE*), has also been suggested as a test termination criterion in the Bayesian adaptive test (Jensema, 1974; Urry, 1977). Therefore, it was of interest to determine how test fairness is influenced under this alternative method of test termination.

One apparent advantage of using s_m is that it would seem to provide a means of reducing differential validity, since all subgroups would simply be tested to the same estimated error level. Since validity bears an inverse relationship to the standard error of estimate (Urry, 1977), all subgroups should attain

nearly equivalent validities. The crucial factor, then, is how well s_m reflects the actual standard error of estimate. For example, if it understates the actual *SEE*, testing will be prematurely terminated, resulting in lower test validity. The results of the present study indicate that this is exactly what happens.

Even when the test items were unbiased, the ratio of the average s_m to the theoretical *SEE* decreased with increasing test length and levels of item discrimination. When item bias was introduced, the ratio decreased with increasing item bias. Consequently, the more biased the test items are, the more likely it is that differential validity will occur when Bayesian adaptive tests are terminated using the observed s_m. The data also suggest that differential validity can be expected to increase the longer the testing process is allowed to continue.

In using the Bayesian adaptive testing strategy to reduce unfairness, it might be possible to compensate for the reduction in validity by differentially setting s_m for each subgroup. Appropriate levels for s_m could be estimated from the data shown in Figures 5 or 6. One problem in devising such a compensatory method is that it could lead to a substantial difference in the average number of test items taken by each subgroup; it has been shown both in the present study and in previous studies that test length affects the other fairness indices.

Advantages of differential prediction. When differential prediction was used, the average s_m became a much better estimator of the theoretical value and was not nearly as adversely influenced by item bias. At α=1.1 it was quite robust with respect to increasing item bias; and even though the underprediction of error increased as item discrimination increased, the use of items with high discriminations is likely to lead to comparable degrees of underprediction for all subgroups. Moreover, with differential prediction, all three of the fairness indices gave convergent implications for the fairness of the adaptive test. Therefore, if Bayesian adaptive testing is terminated on the basis of observed values of s_m, it should be employed within the differential prediction model studied here, in which items are sequentially selected for administration on the basis of subgroup item parameter values.

CONCLUSIONS

All current interpretations of the fairness of a selection procedure depend on the distribution of predicted criterion scores. Both in the present study and in the previous study by Pine and Weiss (1976), it has been shown that the distribution of predicted scores will vary as a function of item characteristics and testing strategy, i.e., adaptive versus conventional. Therefore, even assuming that a selection test is totally free of items biased against a particular subgroup, measured "fairness" will vary as a function of the item characteristics and strategy for selecting items and scoring a test. Thus, the fairness of a test--and consequently a selection program using that test--can vary even if it contains no biased items.

If, in addition, a selection test contains some degree of bias in its items, the situation is compounded. The results of these studies have shown that the extent to which biased items influence selection fairness depends on the testing strategy. Some strategies are more sensitive to the presence of item bias than are others.

In comparing the Bayesian adaptive testing strategy to conventional tests, it was found that the adaptive test was consistently fairer than the conventional tests for tests of 30 or more items with discrimination levels of a=.70 and higher. Furthermore, the differential prediction version of the adaptive test produced almost perfectly fair performance on all fairness indices at high levels of item discrimination. Within the Bayesian strategy it was found that the choice of subgroup prior ability estimates affected test fairness. These effects were minimized by using differential prediction and by increasing test length. Finally, the use of observed values of the Bayesian posterior error of estimate to terminate the Bayesian adaptive test does not assure the reduction of differential validity and can lead to increased unfairness.

One point of caution which needs to be made concerning the use of the Bayesian adaptive testing model is that great care must be taken in choosing an appropriate prior ability distribution for each subgroup. There are essentially two policies which can be followed for each subgroup: (1) using different priors or (2) using identical priors. However, both of these options can result in unfairness--against the minority subgroup in the former case and against the majority subgroup in the latter. Obviously, a dilemma exists. Until this dilemma can be resolved through further research, the use of equivalent prior ability estimates for the majority and minority subgroups is probably advisable, since this will result in the minimum adverse impact on minority subgroups.

Future research efforts on the reduction of test unfairness should be concerned with the problem of prior selection in population subgroups. In addition, other versions of the adaptive approach to testing should be examined to determine their effects on test fairness. There is also a need to determine the kinds and extent of intergroup differences in test item difficulties and discriminations that occur in live-testing situations. These differences should then be incorporated into future simulation studies to determine their interactions with testing strategies and their effects on test fairness. Finally, it is most important that the findings based on theoretical and simulation studies be verified in a live-testing situation.

REFERENCES

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. An investigation of the sources of bias in the prediction of job performance: A six-year study (Research Report PR-73-37). Princeton, NJ: Educational Testing Service, 1973.

Cronbach, L. J., & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17. 127-147.

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 W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper & Row, 1970.

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

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

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

Pine, S. M., & Weiss, D. J. Effects of item characteristics on test fairness (Research Report 76-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Program, December 1976. (NTIS No. AD A035393)

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.

Urry, V. W. A monte carlo study of logistic test models. Unpublished doctoral dissertation, Purdue University, 1969.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Vale, C. D. Strategies of branching through an item pool. In D. J. Weiss (Ed.), Computerized adaptive trait measurement: Problems and prosepects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975. (NTIS No. AD A018675)

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (NTIS No. AD A020961)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

APPENDIX: SUPPLEMENTARY TABLES

Table A
Score Distribution Characteristics for Conventional Tests of Length 10, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	38.4	56.8	45.4	47.4	41.6	43.8	44.0	46.2	44.0	49.8
	.5	min	30.4	46.2	31.8	33.8	28.0	28.2	29.6	30.2	30.6	30.8
		diff	-8.0	-10.6	-13.6	-13.6	-13.6	-15.6	-14.4	-16.0	-13.4	-19.0
	1.0	min	21.6	37.4	20.4	21.4	18.0	17.2	18.2	19.0	17.2	19.2
		diff	-16.8	-19.4	-25.0	-26.0	-23.6	-26.6	-25.8	-27.2	-26.8	-30.6
	2.0	min	10.0	20.2	7.8	7.2	4.4	3.6	4.4	4.6	4.6	4.6
		diff	-28.4	-36.6	-37.6	-40.2	-37.2	-40.2	-39.6	-41.6	-39.4	-45.2
.70												
	0.0	maj	40.6	53.0	50.2	45.0	46.2	49.4	48.0	49.8	48.2	49.2
	.5	min	26.6	35.4	34.4	29.2	28.4	30.4	32.0	29.8	31.0	29.2
		diff	-14.0	-17.6	-15.8	-15.8	-17.8	-19.0	-16.0	-20.0	-17.2	-20.0
	1.0	min	16.4	22.8	19.6	15.0	14.8	15.6	16.2	15.4	15.4	14.8
		diff	-24.2	-30.2	-30.6	-30.0	-31.4	-33.8	-31.8	-34.4	-32.8	-34.4
	2.0	min	4.6	6.2	3.8	3.6	2.8	3.6	2.4	3.4	3.0	3.2
		diff	-36.0	-46.8	-46.4	-41.4	-43.4	-45.8	-45.6	-46.4	-45.2	-46.0
1.1												
	0.0	maj	40.0	52.6	47.2	46.8	47.8	47.6	47.8	48.4	48.8	50.0
	.5	min	25.2	35.2	29.6	29.4	28.0	29.2	27.8	29.2	29.0	29.2
		diff	-14.8	-17.4	-17.6	-17.4	-19.8	-18.4	-20.0	-19.2	-19.8	-20.8
	1.0	min	15.0	20.4	17.4	15.2	15.8	14.6	13.8	15.4	15.4	15.4
		diff	-25.0	-32.2	-29.8	-31.6	-32.0	-33.0	-34.0	-33.0	-33.4	-34.6
	2.0	min	3.4	4.8	3.0	3.4	2.4	3.4	2.2	3.0	2.6	2.6
		diff	-36.6	-47.8	-44.2	-43.4	-45.4	-44.2	-45.6	-45.4	-46.2	-47.4

Table B
Score Distribution Characteristics for Conventional Tests of Length 30, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	38.4	56.8	45.4	47.4	41.6	43.8	44.0	46.2	44.0	49.8
	.5	min	52.6	46.2	40.6	42.6	48.4	47.8	45.0	42.0	46.8	45.4
		diff	14.2	-10.6	-4.8	-4.8	6.8	4.0	1.0	-4.2	2.8	-4.4
	1.0	min	41.8	37.4	49.2	47.8	47.2	44.4	44.4	46.2	46.0	44.2
		diff	3.4	-19.4	3.8	.4	5.6	.6	.4	0.0	2.0	-5.6
	2.0	min	43.6	33.0	41.4	39.0	44.6	45.2	41.4	44.2	44.0	42.4
		diff	5.2	-23.8	-4.0	-8.4	3.0	1.4	-2.6	-2.0	0.0	-7.4
.70												
	0.0	maj	40.6	53.0	50.2	45.0	46.2	49.4	48.0	49.8	48.2	49.2
	.5	min	47.8	47.8	48.2	41.6	47.0	45.8	47.4	46.4	45.8	46.4
		diff	7.2	-5.2	-2.0	-3.4	.8	-3.6	-.6	-3.4	-2.4	-2.8
	1.0	min	49.8	46.0	45.0	41.4	47.0	44.8	45.6	41.8	47.6	42.8
		diff	9.2	-7.0	-5.2	-3.6	.8	-4.6	-2.4	-8.0	-.6	-6.4
	2.0	min	41.2	31.2	44.0	36.6	40.8	38.2	39.4	38.8	42.4	37.4
		diff	.6	-21.8	-6.2	-8.4	-5.4	-11.2	-8.6	-11.0	-5.8	-11.8
1.1												
	0.0	maj	40.0	52.6	47.2	46.8	47.8	47.6	47.8	48.4	48.8	50.0
	.5	min	46.0	43.6	43.8	42.0	45.6	44.0	48.2	43.2	47.4	45.0
		diff	6.0	-9.0	-3.4	-4.8	-2.2	-3.6	.4	-5.2	-1.4	-5.0
	1.0	min	48.4	36.2	48.8	39.6	42.6	39.6	46.6	38.4	45.6	39.4
		diff	8.4	-16.4	1.6	-7.2	-5.2	-8.0	-1.2	-10.0	-3.2	-10.6
	2.0	min	51.0	27.6	40.8	28.6	43.0	30.8	43.0	30.8	44.0	29.6
		diff	11.0	-25.0	-6.4	-18.2	-4.8	-16.8	-4.8	-17.6	-4.8	-20.4

Table C

Values of the C-Index for Uniform (U) and Peaked (P) Conventional Tests and for the Bayesian Adaptive Test (BAT), using Majority and Minority Prediction, and for Majority (maj) and Minority (min) Subgroups, and Subgroup Differences (diff), as a Function of Item Discrimination (a), Degree of Item Bias, for Tests of 10, 30, and 50 Items

			Majority Prediction									Differential Prediction		
			10 Items			30 Items			50 Items			10 Items	30 Items	50 Items
a	Bias	Group	U	P	BAT	U	P	BAT	U	P	BAT	BAT	BAT	BAT
.30		maj	.000	.000	-.024	.000	.000	-.149	.000	.000	-.177	-.024	-.149	-.177
	.5	min	-.124	-.124	-.171	-.258	-.255	-.394	-.317	-.328	-.492	-.045	-.129	-.171
		diff	-.124	-.124	-.147	-.258	-.255	-.245	-.317	-.328	-.315	-.021	.020	.006
	1.0	min	-.254	-.264	-.303	-.527	-.534	-.621	.635	-.664	-.790	-.035	-.111	-.161
		diff	-.254	-.264	-.279	-.527	-.534	-.472	.635	-.664	-.613	.011	.038	.016
	2.0	min	-.509	-.530	-.586	-1.033	-1.023	-1.118	1.262	-1.286	-1.353	-.002	-.097	-.156
		diff	-.509	-.530	-.562	-1.033	-1.023	-.969		-1.286	-1.176	.022	.052	.021
.70		maj	.000	.000	-.044	.000	.000	-.141	.000	.000	-.154	-.044	-.141	-.154
	.5	min	-.283	-.319	-.359	-.377	-.434	-.518	-.422	-.461	-.561	-.074	-.115	-.141
		diff	-.283	-.319	-.315	-.377	-.434	-.377	-.422	-.461	-.407	-.030	.026	.013
	1.0	min	-.586	-.623	-.651	-.801	-.837	-.885	-.879	-.894	-.963	-.073	-.105	-.141
		diff	-.586	-.623	-.607	-.801	-.837	-.744	-.879	-.894	-.809	-.029	.036	.013
	2.0	min	-1.141	-1.142	-1.150	-1.533	-1.524	-1.568	-1.675	-1.634	-1.682	-.070	-.121	-.156
		diff	-1.141	-1.142	-1.106	-1.533	-1.524	-1.427	-1.675	-1.634	-1.528	-.026	.020	-.002
1.1		maj	.000	.000	-.073	.000	.000	-.103	.000	.000	-.119	-.073	-.103	-.119
	.5	min	-.361	-.411	-.434	-.433	-.452	-.509	-.462	-.462	-.541	-.095	-.100	-.121
		diff	-.361	-.411	-.361	-.433	-.452	-.406	-.462	-.462	-.422	-.022	.003	-.002
	1.0	min	-.739	-.780	-.745	-.901	.861	-.896	-.946	-.882	-.946	-.079	-.097	-.123
		diff	-.739	-.780	-.672	-.901	.861	-.793	-.946	-.882	-.827	-.006	.006	-.004
	2.0	min	-1.480	-1.298	-1.227	-1.760	-1.470	-1.538	-1.778	-1.499	-1.599	-.068	-.100	-.121
		diff	-1.480	-1.298	-1.154	-1.760	-1.470	-1.435	-1.778	-1.499	-1.480	.005	.003	-.002
	Priors:													
		maj			.073			-.103			-.119	-.073	-.103	-.119
	-1.0	min			-1.079			-1.097			-1.123	-.303	-.190	-.204
		maj			-1.006			-.994			-1.004	-.230	-.087	-.085
	-.25	min			-.257			-.572			-.661	.255	.104	.054
		maj			-.184			-.469			-.542	.328	.207	.173
	+1.0	min			-.883			-.989			-1.028	-.145	-.122	-.149
		maj			-.810			-.886			-.909	-.072	-.019	-.030

Table D

Values of the T-Index for Uniform (U) and Peaked (P) Conventional Tests and for the Bayesian Adaptive Test (BAT), using Majority and Minority Prediction, and for Majority (maj) and Minority (min) Subgroups, and Subgroup Differences (diff), as a Function of Item Discrimination (a), Degree of Item Bias, for Tests of 10, 30, and 50 Items

			Majority Prediction									Differential Prediction								
			10 Items			30 Items			50 Items			10 Items			30 Items			50 Items		
a	Bias	Group	U	P	BAT	U	P	BAT	U	P	BAT	U	P	BAT	U	P	BAT	U	P	BAT
.30	0.0	maj	38.4	56.8	45.4	45.2	47.4	36.6	41.6	43.8	36.0	38.4	56.8	45.4	45.4	47.4	36.6	41.6	43.8	36.0
	.5	min	30.4	46.2	34.8	31.8	33.8	23.6	28.0	28.2	21.8	52.6	46.2	44.6	40.6	42.6	37.2	48.4	47.8	36.0
		diff	-8.0	-10.6	-10.6	-13.6	-13.6	-13.0	-13.6	-15.6	-14.2	14.2	-10.6	-.8	-4.8	-4.8	.6	6.8	4.0	0.0
	1.0	min	21.6	37.4	26.0	20.4	21.4	15.8	18.0	17.2	12.6	41.8	37.4	45.2	49.2	47.8	38.8	47.2	44.4	35.6
		diff	-16.8	-19.4	-19.4	-25.0	-26.0	-20.8	-23.6	-26.6	-23.4	3.4	-19.4	-.2	3.8	-4.0	2.2	5.6	-.6	-.4
	2.0	min	10.0	20.2	13.0	7.8	7.2	4.6	4.4	3.6	3.6	43.6	33.0	53.6	41.4	39.0	41.8	44.6	45.2	38.0
		diff	-28.4	-36.6	-32.4	-37.6	-40.2	32.0	-37.2	-40.2	-32.4	5.2	-23.8	8.2	-4.0	-8.4	5.2	3.0	1.4	2.0
.70	0.0	maj	40.6	53.0	42.0	50.2	45.0	37.4	46.2	49.4	36.8	40.6	53.0	42.0	50.2	45.0	37.4	46.2	49.4	36.8
	.5	min	26.6	35.4	25.8	34.4	29.2	22.2	28.4	30.4	21.8	47.8	47.8	38.2	48.2	41.6	38.8	47.0	45.8	39.2
		diff	-14.0	-17.6	-16.2	-15.8	-15.8	-15.2	-17.8	-19.0	-15.0	7.2	-5.2	-3.8	-2.0	-3.4	1.4	.8	-3.6	2.4
	1.0	min	16.4	22.8	15.2	19.6	15.0	12.8	14.8	15.6	11.2	49.8	46.0	39.4	45.0	41.4	43.4	47.0	44.8	41.0
		diff	-24.2	-30.2	-26.8	-30.6	-30.0	-24.6	-31.4	-33.8	-25.6	9.2	-7.0	-2.6	-5.2	-3.6	5.0	.8	-4.6	4.2
	2.0	min	4.6	6.2	3.2	3.8	3.6	2.0	2.8	3.6	1.6	41.2	31.2	40.8	44.0	36.6	38.8	40.8	38.2	36.6
		diff	-36.0	-46.8	-38.8	-46.4	-41.4	-35.4	-43.4	-45.8	-35.2	.6	-21.8	-1.2	-6.2	-8.4	1.4	-5.4	-11.2	-.2
1.1	0.0	maj	40.0	52.6	36.4	47.2	46.8	38.6	47.8	47.6	38.8	40.0	52.6	36.4	47.2	46.8	38.6	47.8	47.6	38.8
	.5	min	25.2	35.2	21.6	29.6	29.4	21.0	29.0	29.2	22.4	46.0	43.6	37.2	43.8	42.0	40.6	45.6	44.0	40.0
		diff	-14.8	-17.4	-14.8	-17.6	-17.4	-17.6	-19.8	-18.4	-16.4	5.0	-9.0	.8	-3.4	-4.8	2.0	-2.2	-3.6	1.2
	1.0	min	15.0	20.4	11.8	17.4	15.2	12.0	15.8	4.6	12.0	48.4	36.2	37.8	48.8	39.6	40.4	42.6	39.6	40.0
		diff	-25.0	-32.2	-24.6	-29.8	31.6	-26.6	-32.0	-33.0	-26.8	8.4	-16.4	1.4	1.6	-7.2	1.8	-5.2	-8.0	-.4
	2.0	min	3.4	4.8	1.8	3.0	3.4	1.2	2.4	3.4	1.2	51.0	27.6	39.0	40.8	28.6	39.6	43.0		39.2
		diff	-36.6	-47.8	-34.6	-44.2	-43.4	-37.4	-45.4	-44.2	-37.6	11.0	-25.0	2.6	-6.4	-18.2	1.0	-4.8	16.8	.4
Bayesian Priors		maj			.364			.386			.388			.364			.386			.388
	-1.0	min			.082			.110			.114			.316			.372			.370
		diff			-.282			-.276			-.274			-.048			-.014			-.018
	-.25	min			.202			.138			.138			.540			.468			.448
		diff			-.162			-.248			-.250			.176			.082			.060
	+1.0	min			.106			.118			.116			.380			.408			.382
		diff			-.258			-.268			-.272			.016			.022			-.006