Reduction of Test Bias by Adaptive Testing

Steven M. Pine
University of Minnesota

Because it has the capability of adapting to the specific characteristics of individual testees, computerized adaptive or tailored testing would appear to have potential for reducing test bias due to individual or group difference variables. These variables might include group differences in ability, motivation, test-taking anxiety, or tendency to either guess or omit items.

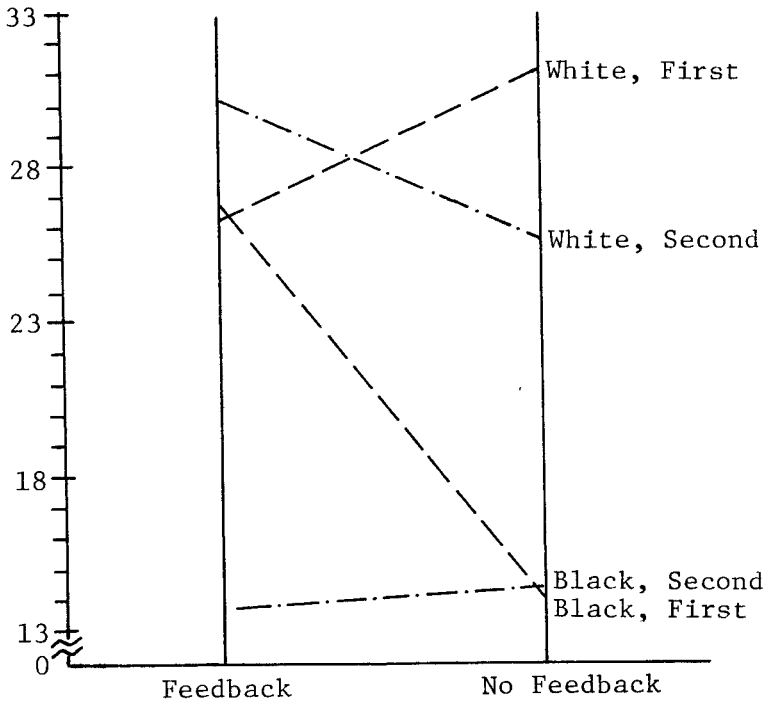
Previous research has provided some evidence for potential psychometric and psychological benefits to minority testees using computerized adaptive testing. Pine and Weiss (in press) demonstrated through a computer simulation that adaptive testing is able to reduce test unfairness and racial differences in test reliability. In a study conducted at the University of Rochester, Johnson and Mihal (1973) administered identical tests to black and white students in a conventional paper-and-pencil format and by computer; however, in this study the computerized test was not adaptive. White students scored significantly higher than blacks on the paper-and-pencil tests, but not on the computer-administered test.

In a second study, conducted at the University of Minnesota, Weiss (1975, p. 24) administered two computer-administered tests to about 100 high school students. The group was racially mixed, consisting of both white and black students. Both a conventional test and a pyramidal adaptive test were administered to each student; half of the group received the conventional test first, and half received the adaptive test first. In addition, half of the group received feedback after each item indicating whether or not their answers were correct; the other half received no feedback after each test item.

The data were analyzed for the conventional test only. Thus, the dependent variable in this analysis was number correct on the conventional test. The design was a $2 \times 2 \times 2$ analysis of variance. The independent variables were (1) race--black and white; (2) feedback--immediate or none; (3) order--conventional test administered first or second in the pair. The results for the three-way analysis of variance showed that the only significant main effect was for race; however, there was a significant three-way order $\times$ race $\times$ feedback interaction.

As shown in Figure 1, when a conventional test was administered first under conditions of immediate feedback, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0)

Figure 1
Mean Scores for Blacks and Whites Completing the 40-Item Conventional Test First and Second, by Feedback Condition



completing the test under the same set of conditions. If this result can be replicated, it implies that race differences observed in test scores may be a function, not of differences in ability, but of differences in psychological effects of the conditions of administration. These findings, although not completely replicating those of Johnson and Mihal (1973), do support their general conclusion that conditions of test administration might affect motivational conditions, which in turn reduce race group differences to nonsignificant levels.

The purpose of the present study was to replicate and extend the previous findings that computerized testing can increase the test scores and the test-taking motivation of minority testees. Specifically, the present study compared a computerized adaptive test designed specifically to minimize test bias with a comparable paper-and-pencil test in order to determine possible racial differences on the following variables:

1. Ability test scores
2. Standard errors of measurement
3. Number of omitted responses
4. Test anxiety, motivation, and tendency to guess.

METHOD

Phase I

The study consisted of two distinct phases. Phase I was intended to develop and calibrate an item pool suitable for testing a group of racially mixed high school students. Three hundred and fifty high school students from two Minneapolis high schools with relatively large minority enrollments were tested. Using these data and data from a previous study, item characteristic curve parameters and an index of bias were calibrated for 250 vocabulary items.

Item Parameters

For each item the Phase I calibration produced estimates for two standard item characteristic curve (ICC) parameters and a third parameter, which was used to index bias. The two ICC parameters estimated were the discriminating power, a, and the item difficulty, b (the guessing parameter, c, was assumed to be equal to .2 for all items). Bias was indexed by an ICC version of Angoff and Ford's (1971) elliptical distance measure of item bias. This index is directly proportional to the difference between the item difficulties of two contrasted subgroups. In the present study this was the difference between the b values for the white and black subgroups.

Table 1
Examples of Vocabulary Item Types

Item Type	Example		
STANDARD	RESCUE	ILLEGAL	FEDERATION
	1. REMEMBER	1. FORBIDDEN	1. RESPECT
	2. REDUCE	2. DISTRESSING	2. ORGANIZATION
	3. MISTAKE	3. ENORMOUS	3. REPORT
	4. SAVE	4. LOYAL	4. GUARANTEE
	5. CHARGE	5. CHEAP	5. INFLATION
BLACK	DOZENS	RANKING	PULPIT
	1. BAKERS	1. MURDERING	1. PREACHING PROFESSION
	2. PERMITS	2. EXCHANGE OF INSULTS	2. ATTRACTIVENESS
	3. FALLS ASLEEP	3. PIG'S INTESTINES	3. QUARRY
	4. INSULTS	4. FRIED COW'S TAIL	4. PLANT TISSUE
	5. DONUTS	5. OLYMPIC EVENT	5. REDUCE
WHITE	BORSCH	TORTE	CAMEO
	1. OVERCOAT	1. CAKE	1. FLOWER
	2. DOG	2. TWIST	2. FRUIT
	3. PORTER	3. SHIRT	3. CRISIS
	4. SOUP	4. CRIME	4. CARVED FIGURE
	5. CHAMBER	5. ANSWER	5. DIAS

Item Types

The item pool consisted of five-alternative multiple-choice vocabulary items of distinct types. About one-half of the items were standard vocabulary

items taken from the University of Minnesota and Educational Testing Service item files. The remaining items, however, were written especially for this study. Of these, thirty were purposely written to be biased against blacks. The remaining items were written specifically for blacks. These items were drawn from three main sources--black literature, two black lexicons, and the items written for this study by a black psychologist. Examples of each of the three item varieties are given in Table 1.

Phase II

In Phase II the calibrated item pool was used to form two standard paper-and-pencil tests and two computerized adaptive tests (CAT) administered on Cathode Ray Tubes (CRT's). Students had ample time to complete the tests and were instructed to guess if they could eliminate at least one alternative as incorrect.

Examinees

Two hundred and thirty students (half white and half black) were tested in Phase II. Of these, the data from 108 blacks and 107 whites were analyzed. Each student was given a McDonald's gift certificate worth $0.50 for participating in the study.

Figure 2
Assignment of Students to Experimental Conditions
for Phase II of Project MINISTEP

Group and Order	FB				NFB			
	BR		NBR		BR		NBR	
	P&P	CAT	P&P	CAT	P&P	CAT	P&P	CAT
Blacks	S_1		S_{15}		S_{29}		S_{43}	
Order 1	⋮		⋮		⋮		⋮	
	S_{14}		S_{28}		S_{42}		S_{56}	
	S_{57}		S_{71}		S_{85}		S_{99}	
Order 2	⋮		⋮		⋮		⋮	
	S_{70}		S_{84}		S_{98}		S_{112}	
Whites	S_{113}		S_{127}		S_{141}		S_{155}	
Order 1	⋮		⋮		⋮		⋮	
	S_{126}		S_{140}		S_{154}		S_{168}	
	S_{169}		S_{183}		S_{197}		S_{211}	
Order 2	⋮		⋮		⋮		⋮	
	S_{182}		S_{196}		S_{210}		S_{224}	

Design

One of the paper-and-pencil tests and one of the computerized tests were designed to minimize test bias. This was referred to as the bias-reduced (BR) test. In addition, half of the students taking each test were given feedback (FB) after each item was administered, indicating whether or not their answer was correct. Each student took one computerized and one paper-and-pencil test. Half of the students took the computerized test first (Order 1), and half took the paper-and-pencil test first (Order 2). In addition, a test reaction questionnaire consisting of motivation, nervousness, guessing, and feedback scales was given to all students at the end of each test condition. Additional dependent measures in this study included three test performance measures indicating test score, standard error of measurement, and number of omitted items. The design of Phase II is shown in Figure 2.

Table 2
Item Characteristics of Items in Stradaptive Test Pool

Stratum	No. Items	Parameter	$\bar{X}$	*S.D.*	Minimum	Maximum
1	11	*a*	.86	.35	.410	1.487
		b	-1.78	.17	-2.039	1.571
		Bias	.74	1.11	-.247	3.730
2	19	*a*	1.00	.34	.531	1.775
		b	-1.12	.22	-1.488	-.909
		Bias	.34	.70	-1.197	.958
3	29	*a*	1.01	.33	.282	1.791
		b	-.58	.18	.889	-.306
		Bias	.53	.59	-.643	1.304
5	43	*a*	1.24	.45	.516	2.268
		b	.02	.18	-.278	.293
		Bias	.37	.51	-.675	1.263
5	35	*a*	1.03	.48	.087	2.215
		b	.59	.20	.315	.884
		Bias	.30	1.30	-5.463	3.526
6	17	*a*	1.13	.43	.447	2.080
		b	1.16	.15	.912	1.460
		Bias	.76	.63	-.519	1.570
7	17	*a*	.78	.70	.900	2.870
		b	3.28	3.60	1.520	16.838
		Bias	.54	3.28	-4.441	10.730

Test Instruments

Computerized adaptive tests. Both of the CAT tests studied--bias-reduced (BR) and non-bias-reduced (NBR)--used the stradaptive strategy developed by Weiss (1973). All items were sorted by difficulty level into one of seven

strata. The two test conditions--BR and NBR--differed only in the way in which items were arranged within each stratum. For the non-biased-reduced (NBR) test conditions, items were arranged by item discrimination level, with the most discriminating items first. For the biased-reduced (BR) test condition, items were arranged by degree of item bias, with the least biased items first. The item characteristics of the stradaptive tests are summarized in Table 2.

An initial stratum assignment was made by asking each testee to rate himself/herself on verbal ability on a three-point scale. The testee was then given the first item in the next-to-easiest, average, or next-to-most difficult stratum depending on his/her self-rating. If the testee's response to this first item was correct, he/she was branched to (administered an item from) the next more difficult stratum. If his/her response was incorrect, he/she was branched to the next easier stratum. If there was not a sufficiently easy or difficult stratum for the required branching (i.e., when an incorrect response was given to an item in Stratum 1, the least difficult stratum, or a correct response given to an item in Stratum 7, the most difficult stratum), the testee was given another item in the same stratum. In all cases the item administered was either the least biased item or the most discriminating (for the NBR condition) item remaining in the stratum.

Paper-and-pencil tests. The bias-reduced (BR) and non bias-reduced (NBR) paper-and-pencil tests were formed from 40 items not used in the stradaptive test item pool. These items were selected--20 from each of the stradaptive test pools--to have approximately the same average item discrimination and item bias as the first 10 items in the BR and NBR stradaptive test pools, respectively. It was impossible to exactly match item characteristics of the 20 items in the paper-and-pencil tests to the 20 used in the stradaptive test, since it could not be determined in advance exactly which 20 items would be administered in the stradaptive tests for each testee. The item characteristics of the paper-and-pencil tests are summarized in Table 3.

Table 3
Item Characteristics of Conventional Paper-and-Pencil Tests

	Bias Reduced			Non-Bias Reduced		
Statistic	*a*	*b*	Bias	*a*	*b*	Bias
Mean	1.57	.05	.80	1.02	.07	-.05
S.D.	.28	.71	.36	.47	.55	1.34
Minimum	1.166	-1.482	.219	.087	-1.482	-5.462
Maximum	2.268	1.136	1.709	2.268	1.009	.744

A special latent ink process was used to give feedback in the paper-and-pencil mode of administration. Students marked their answer sheets with a special pen which caused a latent image (previously invisible) to appear. The letter *Y* appeared if the correct answer was marked; the letter *N* appeared for incorrect answers.

Table 4
Test Reaction Questions for Nervousness and Motivation Scales

Scaled Score	Nervousness Scale
	WERE YOU NERVOUS WHILE TAKING THE TEST?
1	☐ NOT AT ALL
2	☐ SOMEWHAT
3	☐ MODERATELY SO
4	☐ VERY MUCH SO
	DID NERVOUSNESS WHILE TAKING THE TEST PREVENT YOU FROM DOING YOUR BEST?
4	☐ YES, DEFINITELY
3	☐ YES, SOMEWHAT
2	☐ PROBABLY NOT
1	☐ DEFINITELY NOT
	Motivation Scale
	DID YOU CARE HOW WELL YOU DID ON THE TEST?
4	☐ I CARED A LOT
3.2	☐ I CARED SOME
2.4	☐ I CARED A LITTLE
1.6	☐ I CARED VERY LITTLE
.8	☐ I DIDN'T CARE AT ALL
	DID YOU FEEL CHALLENGED TO DO AS WELL AS YOU COULD ON THE TEST?
1	☐ NOT AT ALL
2	☐ SOMEWHAT
3	☐ FAIRLY MUCH SO
4	☐ VERY MUCH SO
	WERE YOU INTERESTED IN KNOWING WHETHER YOUR ANSWERS WERE RIGHT OR WRONG?
4	☐ I WAS VERY INTERESTED
3	☐ I WAS MODERATELY INTERESTED
2	☐ I WAS SOMEWHAT INTERESTED
1	☐ I DIDN'T CARE AT ALL

Table 5
Test Reactions Questions for Guessing and Feedback Scales

Scaled Score	Guessing Scale
	ON HOW MANY OF THE QUESTIONS DID YOU GUESS?
4	☐ ALMOST ALL OF THE QUESTIONS
3.33	☐ MORE THAN HALF OF THE QUESTIONS
2.67	☐ ABOUT HALF OF THE QUESTIONS
2	☐ LESS THAN HALF OF THE QUESTIONS
1.33	☐ ALMOST NONE OF THE QUESTIONS
.67	☐ NONE OF THE QUESTIONS
	HOW OFTEN WERE YOU SURE THAT YOUR ANSWERS TO THE QUESTIONS WERE CORRECT?
.8	☐ ALMOST ALWAYS
1.6	☐ MORE THAN HALF OF THE TIME
2.4	☐ ABOUT HALF OF THE TIME
3.2	☐ LESS THAN HALF OF THE TIME
4	☐ 5. ALMOST NEVER
	Feedback Scale
	DID RECEIVING FEEDBACK AFTER EACH QUESTION INTERFERE WITH YOUR ABILITY TO CONCENTRATE ON THE TEST?
1	☐ NO, NOT AT ALL
2	☐ YES, SOMEWHAT
3	☐ YES, MODERATELY SO
4	☐ YES, VERY MUCH SO
	DID GETTING FEEDBACK AFTER EACH QUESTION MAKE YOU NERVOUS?
1	☐ NO, NOT AT ALL
2	☐ YES, SOMEWHAT
3	☐ YES, MODERATELY SO
4	☐ YES, VERY MUCH SO

Test reaction questions. The psychological reactions to each testing condition were assessed by administering test reaction questions consisting of four factor-analytically-derived scales designed to measure (1) nervousness, (2) motivation, (3) tendency to guess, and (4) reaction to feedback. (See Prestwood & Weiss, 1977, for a description of the derivation of these scales.) Tables 4 and 5 show the test reaction items by scale. Items were administered to each testee twice--once after each testing condition. Testees in the no feedback condition were given only the motivation, nervousness, and guessing scales.

Test Performance Measures

Three test performance measures, indicating ability test score, standard error of measurement, and number of omitted responses were investigated. The ability test score was obtained by a Bayesian scoring procedure similar to the one developed by Owen (1975). This procedure provided a means of generating comparable scores for both the conventional and adaptive tests. The posterior Bayesian variance was used as an estimate of the standard error of measurement; Jensema (1974) has indicated the relationship between these measures.

RESULTS AND DISCUSSION

Test Performance Measures

Significant F ratios for the 2×2×2×2×2 repeated measures analysis of variance performed on the Bayesian ability scores, the Bayesian variance, and the number of omitted responses are given in Table 6.

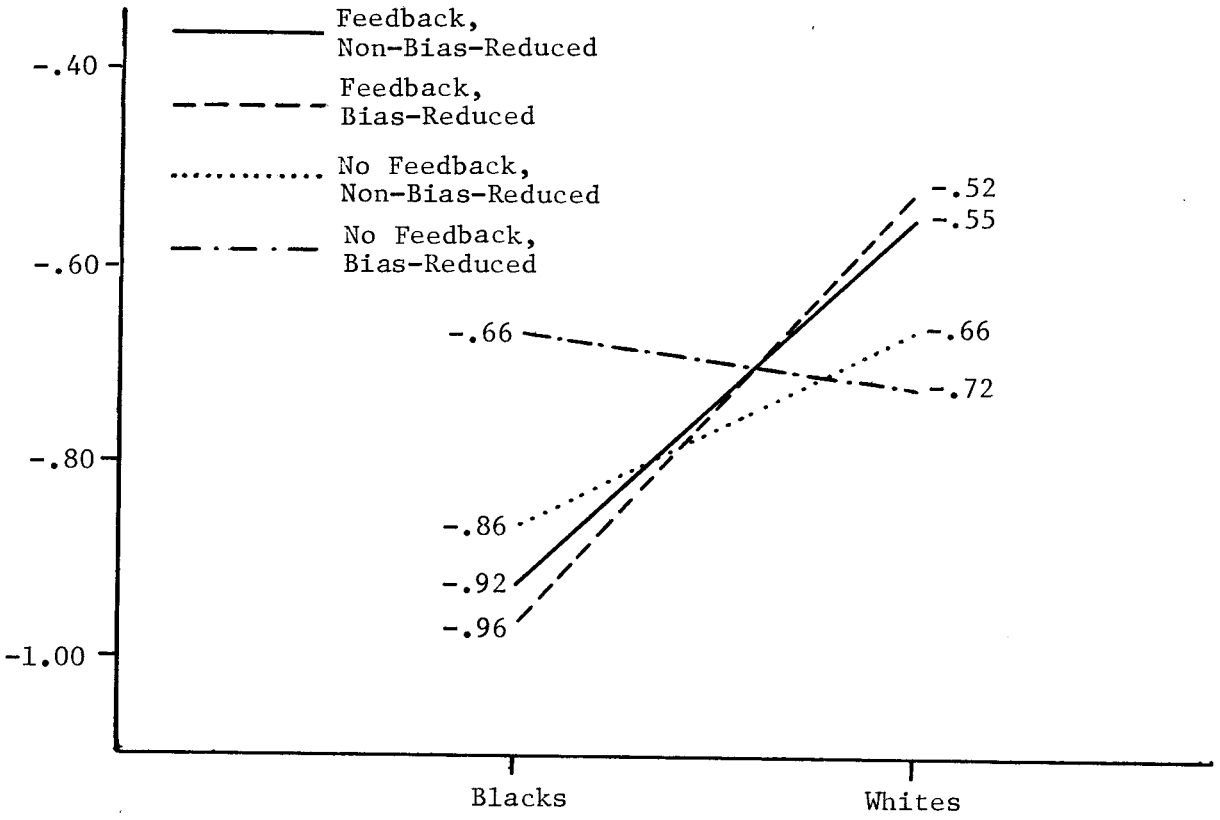
Table 6
Significant F Ratios for ANOVAs on Performance Measures

Dependent Variable	Source	Degrees of Freedom	Mean Square	F	p
Bayesian Score	R	1	5.83	6.60	.011
	RFB	1	2.93	3.31	.070
	Error (R,RFB)	199	.88		
	MRFB	1	.56	3.65	.057
	Error (MRFB)	199	.53		
Bayesian Variance	B	1	.46	582.28	.000
	ROB	1	.00	4.69	.031
	Error (B,ROB)	199	.01		
	MO	1	.02	13.17	.000
	MB	1	.01	6.01	.015
	MOB	1	.01	9.32	.003
	Error (MO,MB,MOB)	199	.01		
Number of Omits	F	1	96.27	4.43	.037
	RF	1	90.10	4.15	.043
	Error (F,RF)	199	21.72		

Bayesian ability estimates. The only significant ($p<.02$) main effect on the Bayesian ability scores was for race (R), with whites scoring significantly higher than blacks. This is the result commonly found in the testing literature. It is, however, particularly noteworthy in the present study, since one of the test conditions (BR) was specifically designed to reduce test score differences and included "Black-type" items. This result cannot alone be interpreted as indicating that the BR condition was ineffective, since the size of this effect may have been larger had the BR condition not been included.

Racial effects resulting from the manipulation of the bias-reduction (B), feedback (F), and mode (M) variables can be interpreted in the significant interactions involving the race variable. Two such interactions were significant. The race × feedback × bias-reduced interaction, which is shown in Figure 3, is particularly interesting.

Figure 3
Race × Bias-Reduced × Feedback Interaction for Bayesian Scores



The influence of feedback was straightforward for whites. Whites performed better on both tests (BR and NBR) when feedback was provided. The results were different for blacks. Blacks performed worse on both tests when feedback was given; but when feedback was not given, they scored better on the bias-reduced test.

These results may reflect race differences in the rate and aversiveness of negative feedback. Blacks received more negative feedback than did whites

because they tended to answer more items incorrectly. Furthermore, the negative feedback which blacks received in the bias-reduced condition was likely to be particularly aversive, since they were being told they did not know the meaning of "black-type" words. The four-way race × bias-reduced × feedback × mode interaction indicated that although blacks scored lower when feedback was given, they did relatively better when the items were administered by computer with the adaptive testing strategy.

Bayesian posterior variance. The analysis of variance performed on the Bayesian posterior variance dependent measure produced a significant main effect for the type of test (BR or NBR) and for the race × order × bias-reduced and mode × order × bias-reduced interactions. It should be recalled that (1) the Bayesian posterior variance provides an estimate of the standard error of measurement and (2) the only distinction between the BR and NBR stradaptive tests was the order in which the items were arranged within each stratum. In the BR tests, items were arranged from least biased to most biased; in the NBR tests, items were in descending order of their item discriminations. Therefore, these results are consistent with the fact that the standard error of measurement at a fixed ability level is decreased by increasing item discrimination. The mode × bias-reduced interaction provided additional support to the growing body of research (Vale & Weiss, 1975a, 1975b; Betz & Weiss, 1976a, 1976b), showing that stradaptive tests produce smaller standard errors of measurement than comparable conventional tests.

Number of omits. The administration of feedback significantly ($p<.05$) reduced the number of omitted items. The significant two-way interaction between race and feedback shown in Figure 4 indicates the nature of the feedback effect.

Figure 4
Mean Number of Omits as a Function
of Feedback for Blacks and Whites

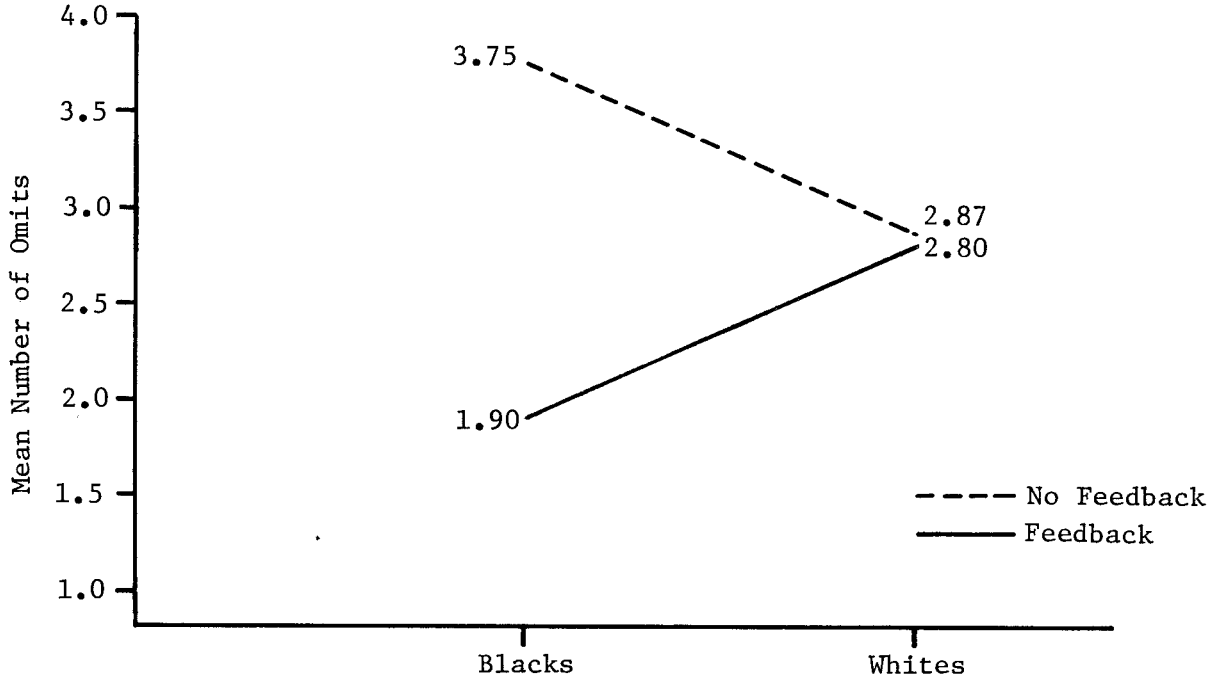


Figure 4 indicates that the significance of the feedback effect was almost entirely attributable to its influence on blacks. When feedback was administered, blacks omitted significantly ($p<.05$) fewer items. However, this reduction in the tendency to omit items did not lead to a significant increase in test scores for blacks. This is probably because a decrease in the number of omits was accompanied by an increase in guessing, which would not be expected to significantly increase a Bayesian ability estimate.

Test Reaction Scales

Table 7 gives the results of the analysis of variance for the test reaction scales. Significant F ratios were obtained for the feedback, nervousness, motivation, and tendency-to-guess scales.

Table 7
Significant F Ratios for ANOVAs on Test Reaction Scales

Dependent Variable	Source	Degrees of Freedom	Mean Square	F	p
Feedback	R	1	7.78	8.60	.00
	Error (R)	80	.91		
	MO	1	.87	4.13	.05
	Error (MO)	80	.21		
Nervousness	OB	1	5.37	7.46	.01
	ROFB	1	3.17	4.41	.04
	Error (OB,ROFB)	185	.72		
	M	1	1.22	5.01	.03
	MB	1	.87	3.57	.06
	Error (M,MB)	185	.24		
Motivation	RO	1	4.68	5.41	.02
	ROFB	1	5.10	5.89	.02
	Error (RO,ROFB)	185	.87		
	M	1	2.17	14.04	.00
	MB	1	1.25	8.09	.01
	MROB	1	.83	5.35	.02
	Error (M,MB,MROB)	185	.15		
Guessing	OFB	1	2.74	3.67	.06
	Error (OFB)	185	.75		
	M	1	2.06	6.06	.02
	MRO	1	1.26	3.72	.06
	MOB	1	1.53	4.51	.04
	Error (M,MRO,MOB)	185	.34		

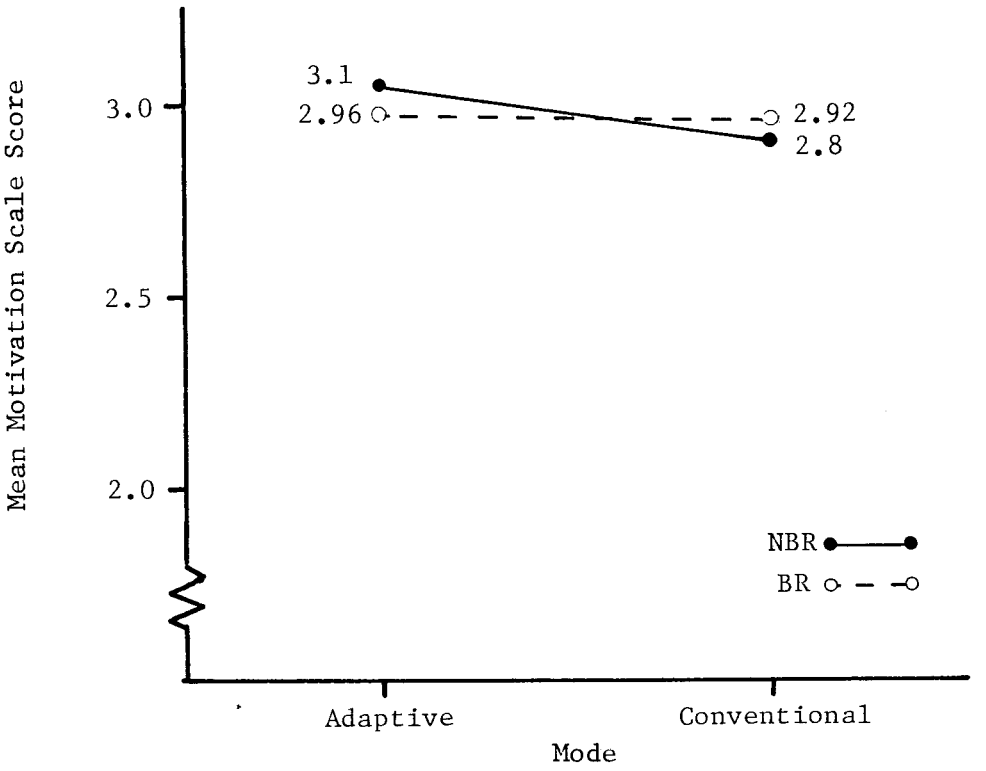
Feedback scale. The significant ($p<.005$) race (R) effect for the feedback scale indicated that blacks were more adverse to receiving feedback than were whites. They felt that receiving feedback impaired their ability to concentrate and made them somewhat nervous. This finding is consistent with the negative feedback hypothesis and may, at least partially, explain why blacks scored lower than whites.

Nervousness scale. The only significant ($p<.05$) main effect for the nervousness scale indicated that testees, in general, were more nervous when

tested by an adaptive testing strategy than when tested conventionally (M effect). The only significant effect ($p<.05$) involving race was the four-way race × order × feedback × bias-reduced interaction (ROFB). Because of the number of terms involved, it is difficult to interpret the exact meaning of this interaction. One explanation of the main effect, however, can be attributed to the tendency of many people to become nervous when required to work in a novel environment, particularly when that environment involves the operation of a machine. Another explanation can be attributed to the nature of the adaptive testing procedure. In adaptive testing, item difficulties are tailored to the ability level of each testee. Therefore, testees are constantly being challenged to the limit of their ability. This would understandably increase nervousness or test anxiety.

Motivation scale. Again, the only significant ($p<.001$) main effect was due to mode of test administration (M), indicating higher test-taking motivation in the adaptive testing condition. The significant ($p<.005$) mode × bias-reduced (MB) interaction (see Figure 5) shows that motivation did not vary as a function of the biasedness of the test in the adaptive testing condition, but did in the paper-and-pencil condition. These results support results in the adaptive testing literature (Weiss & Betz, 1973; Weiss, 1974) that test-taking motivation can be increased with adaptive testing.

Figure 5
Mean Motivation Scale Scores as a Function of Mode of Administration and Bias-Reduction



Significant race effects emerged in the race × order (RO), race × order × feedback × bias-reduced (ROFB), and mode × race × order × bias-reduced (MROB) interactions. In conjunction with the fact that the race × order × feedback × bias-reduced interaction was also found to be significant on the nervousness scale, there appears to be sufficient evidence to indicate a racial difference in the psychological aspects of test-taking behavior.

Guessing scale. Once again, the only significant ($p<.02$) main effect was for mode of administration (M), with a lesser tendency to guess in the adaptive testing condition. This finding is consistent with the fact that adaptive tests tend to administer fewer items above a testee's ability level; consequently, there is less reason for an individual to guess.

Summary and Conclusions

The present study provided general information on the relative merits of computerized adaptive versus conventional paper-and-pencil testing and provided specific information on the use of these two methods for testing minority examinees. The results of the study add to the growing body of research which has shown the general superiority of adaptive testing over conventional paper-and-pencil testing. It was found that in comparison to paper-and-pencil tests, computerized stradaptive tests (1) produced ability estimates with smaller standard errors of measurement, (2) reduced the tendency to guess, and (3) increased test-taking motivation.

Evidence also emerged which indicated that people are anxious (nervous) following an adaptive test. This finding is understandable, since in adaptive testing, items are tailored to the ability level of each testee and will therefore tend to challenge the testee to the limits of his/her ability. This increased nervousness did not, however, decrease performance or reduce the accuracy of test scores.

No definitive evidence was manifested to support the claims that either feedback and/or adaptive testing significantly improves the test performance of blacks; nor did the bias-reducing condition produce a significant increase in black test scores. Several important factors about the conditions of this study, however, must be kept in mind when interpreting these results. First, of the 20 items administered to each student, only about 5 were likely to be words more common in black culture. In view of the relatively small cell frequencies (between 12 and 15), it is unlikely that there was sufficient power to detect a test score increase for blacks attributable to approximately five items. Furthermore, blacks did score nonsignificantly higher overall on the bias-reduced tests and on the adaptive bias-reduced tests, thereby suggesting that the conditions provided by these tests may have been beneficial to minority testees.

Secondly, in order to evaluate the extent to which these results support the feedback effect reported by Weiss (1975, p. 35), several important differences between these studies must be considered. The Weiss study did not include a bias-reduced condition and gave feedback which was specifically chosen to be meaningful to blacks. (The feedback in the present study con-

sisted of a simple "correct", or "incorrect.") In those conditions of the present study corresponding most closely to the conditions of the Weiss study, a similar feedback effect was found--blacks scored relatively higher ($p<.15$) when they were given feedback on the computer. Furthermore, test-taking motivation was significantly higher for blacks overall and was non-significantly higher for blacks in the computerized feedback condition. Also, blacks omitted significantly fewer items than did whites when feedback was given.

The apparently contradictory finding that blacks scored significantly lower when given feedback on the computer occurred only in the bias-reduced condition. Blacks may have been particularly distressed upon being told that they didn't know the meaning of words common to their culture. This hypothesis is supported by blacks reporting that getting feedback made them significantly more nervous and somewhat reduced their ability to concentrate.

The finding from the test reaction scale concurs with the general conclusions of the Johnson and Mihal (1973) and Weiss (1975) studies in demonstrating that blacks react differently to the conditions of testing (e.g., feedback and mode of test administration) than do whites. Although race test score differences were not reduced to nonsignificant levels in the present study, the direction of test score differences was consistent with the conclusions of previous studies. This suggests that racial difference in test scores can be reduced with computerized testing techniques.

References

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude (Research Bulletin RE-71-59). Princeton, NJ: Educational Testing Service, 1971.

Betz, N. E., & Weiss, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027147) (a)

Betz, N. E., & Weiss, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing (Research Report 76-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027170) (b)

Jensema, C. J. An application of latent trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 20-48.

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

Owen, R. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Pine, S. M., & Weiss, D. J. A comparison of the fairness of adaptive and conventional testing strategies (Research Report 78-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, in press.

Prestwood, J. S., & Weiss, D. J. Accuracy of perceived test-item difficulties (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977 (NTIS No. AD A041084)

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (NTIS No. AD A018758) (a)

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) (b)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

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. Adaptive testing research at Minnesota: Overview, recent results, and further directions. In C. L. Clark (Ed.), Proceedings of the First Conference on Computerized Adaptive Testing. Washington, DC: U.S. Civil Service Commission, 1976.

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)

Acknowledgments

This research was supported by contract N00014-76-C-0244, NR No. 150-383, with the Personnel and Training Research Programs, Office of Naval Research. The author is indebted to Frederick Lord of the Educational Testing Service for providing many of these items and also wishes to thank Harvey Linder and Ken Jones for writing many of the items used in this study.