PROSPECTS: NEW TYPES OF INFORMATION AND PSYCHOLOGICAL IMPLICATIONS

Nancy E. Betz
University of Minnesota

Traditional psychometric theory and practice has largely failed to take advantage of the full variety and extent of information obtainable from responses to test items. Consequently, the most information usually extracted from a testee's responses to a series of items is a total test number correct score, or a score on some personality dimension or interest scale.

But patterns of test item responses are far richer in information and are far more complex to interpret than single number correct scores would imply. Computer-assisted testing procedures provide us with the capability of extracting much more and a greater variety of information about an individual or about the meaning of his/her score than have conventional testing procedures.

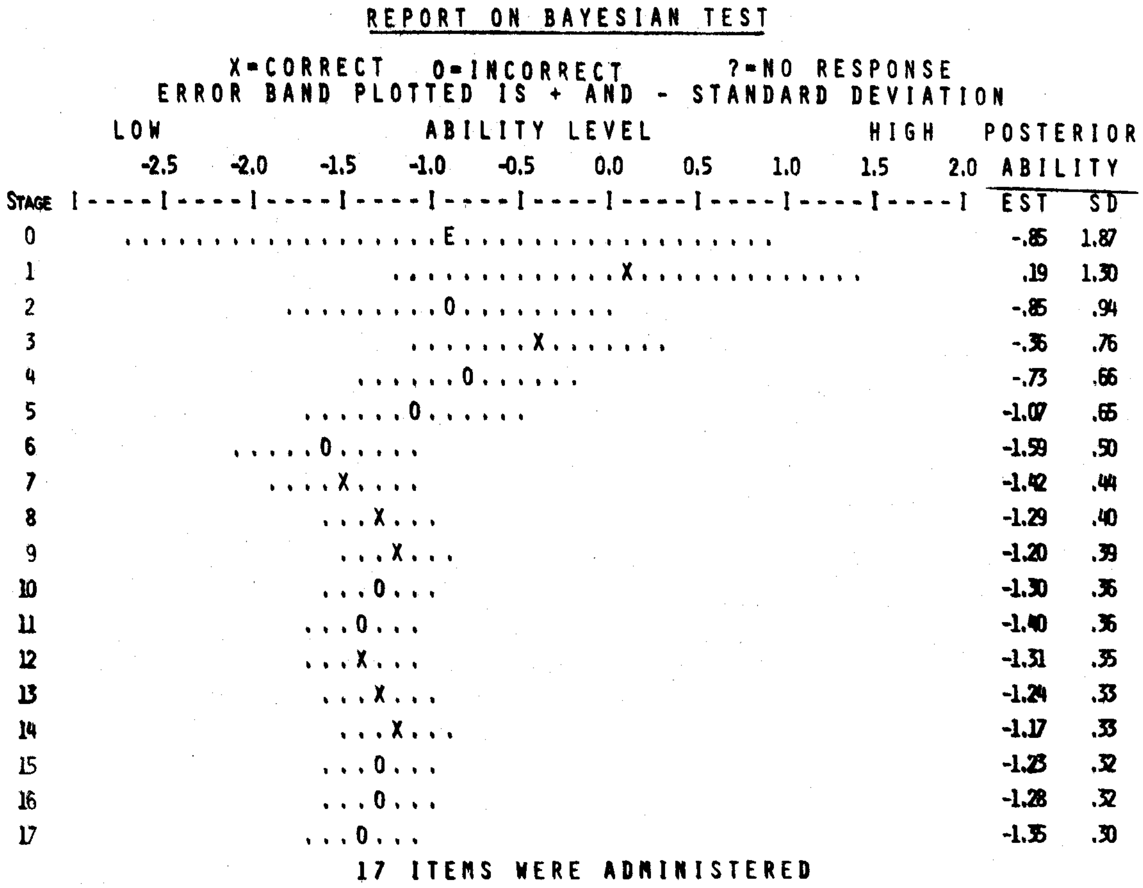
New Types of Information

Individualized errors of measurement. Probably one of the most important new types of information obtainable from computerized adaptive trait measurement procedures is a value indicating the accuracy of a given individual's score on a test--that is, a value indicating the degree of confidence we can place in a particular individual's test score. The traditional psychometric approach to this problem has involved the determination of a reliability coefficient characterizing a whole test--from that reliability coefficient we derive a standard error of measurement which we use to estimate the amount of probable error in a given individual's test score. However, this standard error of measurement represents the average expected error over all individuals in the group and, as Mr. Vale has shown, the error in a typical peaked conventional test is much greater for individuals whose ability levels deviate from the average. Consequently, the *average* expected error may be an overestimate or an underestimate of the amount of error in any one score.

Several of the adaptive testing strategies provide individualized estimates of score accuracy. For example, the Bayesian adaptive testing strategy provides, along with an ability estimate following each item administered, a value indicating the error of that estimate.

Figure 24 shows an example of ability estimates and errors obtained as successive items are administered to an individual in a Bayesian adaptive test. Note how the size of the error band around the ability estimate decreases as responses to successive items provide us with more information and a more stable estimate of ability.

Figure 24



```
                              REPORT ON BAYESIAN TEST

                    X=CORRECT   O=INCORRECT      ?=NO RESPONSE
                ERROR BAND PLOTTED IS + AND - STANDARD DEVIATION
        LOW                ABILITY LEVEL                 HIGH   POSTERIOR
          -2.5 -2.0 -1.5 -1.0 -0.5  0.0  0.5  1.0  1.5  2.0  ABILITY
STAGE I----I----I----I----I----I----I----I----I----I----I     EST    SD
  0      ..................E..................                -.85  1.87
  1                     .............X.............            .19  1.30
  2               .........0.........                         -.85   .94
  3                      .......X.......                      -.36   .76
  4                   ......0......                           -.73   .66
  5                ......0......                             -1.07   .65
  6            .....0.....                                   -1.59   .50
  7              ....X....                                   -1.42   .44
  8                 ...X...                                  -1.29   .40
  9                  ...X...                                 -1.20   .39
 10                 ...0...                                  -1.30   .36
 11                ...0...                                   -1.40   .36
 12                ...X...                                   -1.31   .35
 13                 ...X...                                  -1.24   .33
 14                  ...X...                                 -1.17   .33
 15                 ...0...                                  -1.23   .32
 16                 ...0...                                  -1.28   .32
 17                ...0...                                   -1.35   .30
                         17 ITEMS WERE ADMINISTERED
```

In Bayesian adaptive testing, we can either fix the number of items administered, thus allowing the error of the ability estimate to vary across individuals, or we can administer different numbers of items to different individuals with the intention of terminating the test when an acceptably small degree of measurement error has been achieved. Thus, the Bayesian ability estimate is far more interpretable than are conventional test scores because we can obtain an estimate of the amount of probable error in each individual's score.

Response consistency. Another type of information obtainable from some adaptive testing strategies is something that we have called the *consistency* of an individual's response pattern. Consistency refers to how reliably or consistently an individual is interacting with an item pool.

In personality assessment, response inconsistency is usually assessed using various types of validity scores. The notion of inconsistency in, for example, pair comparisons or forced choice formats, is operationalized as the number of circular triads. If a person's response pattern contains too many circular triads, we infer that something besides the trait of interest is influencing the person's responses and declare his test protocol invalid.

In ability measurement, we would expect that an individual should, in general, respond correctly to items below, or easier than, his/her ability level, and incorrectly to items above, or more difficult than, his/her ability level. If a person answers most easy items correctly and most difficult items incorrectly, we would say that he is responding consistently--that is, his response pattern seems to be influenced primarily by his position on the underlying trait continuum. However, if a person answers many easy items incorrectly and many difficult items correctly, he is responding inconsistently, indicating that something besides the trait of interest is influencing his responses.

In an ability test, response inconsistency may be caused by such extraneous variables as guessing, partial knowledge, or adverse psychological conditions such as test anxiety or lack of motivation to do one's best on the test. Whatever its cause, response inconsistency may reduce the reliability and/or validity of a given test score. And, knowing the degree of consistency of an individual's response pattern may be important when we intend to use that score in making practical decisions.

We have operationalized the notion of response consistency in the stradaptive testing strategy. As you may recall from Mr. Vale's presentation (see Figure 15), in the stradaptive test, items are organized into a series of levels or strata according to their difficulty. A correct response to an item in one stratum leads to the administration of the most discriminating item remaining in the next more difficult stratum. An incorrect response leads to the administration of the most discriminating item remaining in the next less difficult stratum.

Figure 25 shows a relatively consistent response pattern on the stradaptive test along with 10 ability scores and five consistency scores. This person entered the stradaptive test at stratum 5, based on some prior information. Stratum 5 items were too easy for him and he answered items correctly until, at item 4, he had been branched to stratum 8, which contained very difficult items. Notice that he consistently responded incorrectly to the stratum 8 items, which were too difficult for him, and correctly to the stratum 6 items, which were too easy for him. The items in stratum 7 seem most appropriate in difficulty, and he answered about half of them correctly and the other half incorrectly.

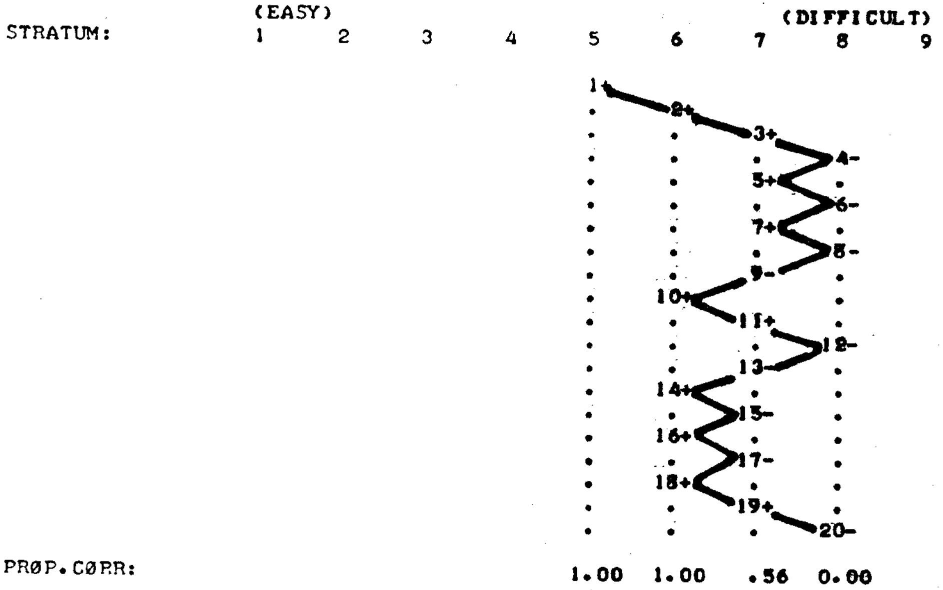
The consistency of this individual's response pattern was reflected in his relatively low consistency scores. Score 11, defined as the standard deviation of the difficulties of the items encountered by this person, was .59. Further, in the stradaptive test, items are administered until a termination criterion is reached. The consistency of this individual's response pattern enabled him to meet the termination criterion after only twenty items had been administered.

Contrast the response pattern of this consistent examinee with the one shown in Figure 26. The response pattern shown in Figure 26 was far less consistent and ranged over a larger number of strata, and thus a larger range of item difficulty. For example, this person answered some relatively easy items at stratum 5 incorrectly (e.g., items 8 and 26) and answered some difficult items at stratum 8 correctly (e.g., items 1 and 17). By responding inconsistently, it took many more items before the termination criterion was reached, and the individual's consistency scores are higher, reflecting a less consistent response pattern.

Figure 25

Report on a Stradaptive Test for a Consistent Testee

```
                    REPØRT ØN STRADAPTIVE TEST
ID NUMBER:                                  DATE TESTED:  73/07/12
------------------------------------------------------------------
```



```
TØTAL PRØPØRTIØN CØRRECT=  .550
```

SCØRES ØN STRADAPTIVE TEST

Ability Level Scores

```
 1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT=  1.49

 2. DIFFICULTY ØF THE N+1 TH ITEM=  1.44

 3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT=  1.49

 4. DIFFICULTY ØF HIGHEST STRATUM
    WITH A CØRRECT ANSWER=  1.33

 5. DIFFICULTY ØF THE N+1 TH STRATUM=  1.33

 6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM=  1.33

 7. INTERPØLATED STRATUM DIFFICULTY=  1.37

 8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS=   .88

 9. MEAN DIFFICULTY ØF CØRRECT ITEMS
    BETWEEN CEILING AND BASAL STRATA=  1.28

10. MEAN DIFFICULTY ØF ITEMS CØRRECT
    AT HIGHEST NØN-CHANCE STRATUM=  1.28
```

Consistency Scores

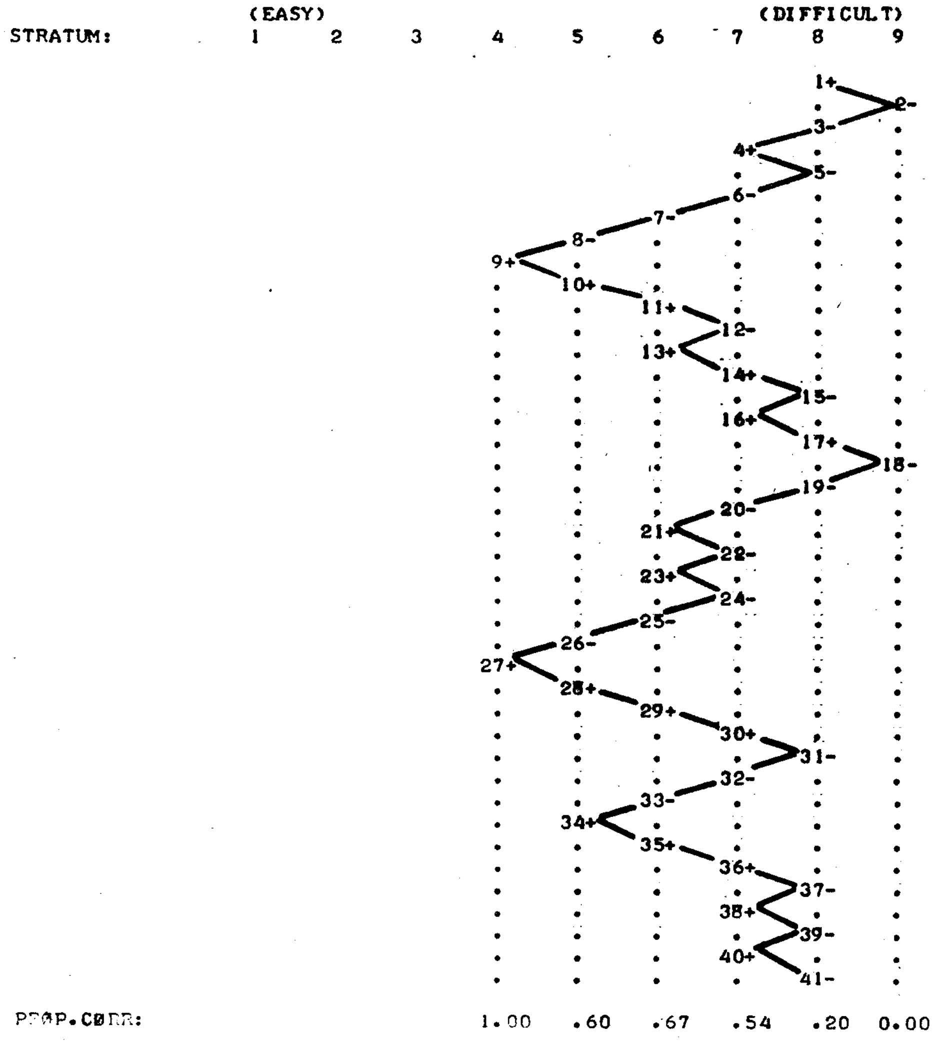
```
11. SD ØF ITEM DIFFICULTIES ENCØUNTERED=   .59

12. SD ØF DIFFICULTIES ØF
    ITEMS ANSWERED CØRRECTLY=   .46

13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED
    CØRRECTLY BETWEEN CEILING AND BASAL STRATA=   .18

14. DIFFERENCE IN DIFFICULTIES
    BETWEEN CEILING AND BASAL STRATA=  1.36

15. NUMBER ØF STRATA BETWEEN
    CEILING AND BASAL STRATA=  1
```

Figure 26

Report on a Stradaptive Test for an Inconsistent Testee

REPØRT ØN STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/02



TØTAL PRØPØRTIØN CØRRECT= .488

SCØRES ØN STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY ØF MØST DIFFICULT ITEM CØRRECT= 1.89
2. DIFFICULTY ØF THE N+1 TH ITEM= 1.01
3. DIFFICULTY ØF HIGHEST NØN-CHANCE ITEM CØRRECT= 1.53
4. DIFFICULTY ØF HIGHEST STRATUM WITH A CØRRECT ANSWER= 2.01
5. DIFFICULTY ØF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY ØF HIGHEST NØN-CHANCE STRATUM= 1.33
7. INTERPØLATED STRATUM DIFFICULTY= 1.36
8. MEAN DIFFICULTY ØF ALL CØRRECT ITEMS= .72
9. MEAN DIFFICULTY ØF CØRRECT ITEMS BETWEEN CEILING AND BASAL STRATA= .76
10. MEAN DIFFICULTY ØF ITEMS CØRRECT AT HIGHEST NØN-CHANCE STRATUM= 1.24

Consistency Scores

11. SD ØF ITEM DIFFICULTIES ENCØUNTERED= .86
12. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY= .74
13. SD ØF DIFFICULTIES ØF ITEMS ANSWERED CØRRECTLY BETWEEN CEILING AND BASAL STRATA= .50
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 2.64
15. NUMBER ØF STRATA BETWEEN CEILING AND BASAL STRATA= 3

Consistency and stability. We hypothesized that the ability test scores of individuals who are responding consistently should be more reliable than those of individuals who are responding inconsistently. To study this hypothesis, we used test-retest stability as an indication of score reliability, and divided a group of 200 subjects into five groups on the basis of their consistency scores on the first stradaptive test administration in a test-retest design. Within each group, we calculated the test-retest stability of the obtained ability estimates. Table 1 shows the results obtained for the consistency score defined as the standard deviation of the difficulties of all items encountered.

Table 1

STRADAPTIVE AND CONVENTIONAL TEST TEST-RETEST CORRELATIONS AS A FUNCTION OF CONSISTENCY SCORE 11 ON INITIAL TESTING

		STATUS ON CONSISTENCY SCORE 11				
		VERY HIGH	HIGH	AVERAGE	LOW	VERY LOW
MEAN CONSISTENCY SCORE		.517	.625	.706	.815	1.038
NUMBER OF TESTEES IN INTERVAL		27	30	41	43	29
STRADAPTIVE ABILITY SCORE:	1	.940	.849	.847	.768	.652
	2	.875	.721	.799	.778	.751
	3	.956	.813	.878	.826	.708
	4	.934	.840	.846	.731	.664
	5	.896	.722	.793	.756	.741
	6	.950	.798	.886	.820	.704
	7	.970	.844	.902	.851	.758
	8	.981	.927	.915	.853	.869
	9	.983	.939	.907	.899	.889
	10	.951	.792	.882	.822	.718
CONVENTIONAL TEST		.979	.890	.918	.826	.878

As the table shows, the highest test-retest stability was found in the most consistent group of examinees for all ten ability scores. The clearest pattern is that for ability score 1, where the scores in the most consistent group had a test-retest stability of .94, while the scores in the least consistent group had a stability of .65. The stabilities in the intermediate groups decreased with decreasing consistency. Note also that the stability for the most consistent examinees on scores 8 and 9 was .98, indicating very high stability of the obtained ability estimates. These results suggest that the use of consistency scores as moderator variables may provide us with additional information concerning the accuracy of longitudinal predictions from test scores.

Thus, such indices as estimates of the degree of accuracy of a given individual's test score or the consistency of a test response protocol add greatly to our capacity to meaningfully interpret a test score, and to the utility that the score will have in practical decision-making contexts.

Additional new kinds of information. Computerized trait measurement can provide us with additional types of information. For example, the computer can provide precise control over a subject's usage of confidence weighting procedures or probabilistic responding, which can be used to assess partial knowledge. When confidence weighting has been used in a paper and pencil format, it has frequently been found that some examinees fail to assign probabilities to the response alternatives in accordance with the test instructions. In computer-administration, however, the examinee is informed immediately when he has not assigned probabilities according to the rule. Thus computerized test administration can eliminate the problem of invalid test protocols.

Computerized testing also has the capability of providing us with exact response latency data for each item administered. Response latency data have a variety of potential uses. For example, it might be used in conjunction with confidence weighting procedures to aid in the identification of guessing behavior. In the area of personality assessment it could be useful in identifying the presence of random responding or response sets. Finally, the measurement of response latencies may lead to further understanding of the speed versus power components of ability.

Perhaps the most potentially important and fruitful area of research using computerized testing lies in the study of human problem-solving and reasoning abilities. Traditionally, psychometricians have asked *how many* problems a person could solve and have left it to the experimentalists to investigate the nature or the "how" of the problem-solving process. But knowledge of the process of problem-solving should be a part of our theories of human abilities and could contribute substantially to the construct validity of such theories.

One approach to the study of problem-solving abilities using computerized test administration would involve a within-problem branching sequence in which a series of interdependent questions are organized into a problem-oriented structure. For example, one response at a given point in the structure might result in the testee's arriving at a correct solution by an entirely different pathway than would a different response at that given point. We could study the amount and type of information the testee needs to solve a problem, the efficiency with which he goes about it, and the different problem-solving systems or pathways utilized by different individuals.

The time now seems right, therefore, for using the computer to integrate the measurement and the study of intelligent behavior. Limiting the information we obtain from test-taking behavior to whether an item was answered correctly or incorrectly is wasteful of much potentially significant and useful information and is now no longer necessary, thanks to the availability of computer-assisted testing procedures.

Psychological Effects

In addition to the variety of new information obtainable from computer-assisted testing procedures, it also has the potential to improve the psychological environment of testing. In the past, psychometricians have paid considerable attention to the characteristics of tests administered to groups, for example, their reliability and validity. But we have forgotten that it is an *individual* who takes a test, not a group. Highly valid and reliable tests can be rendered useless for the measurement of an individual if, for one reason or another, he is not performing to his fullest capacity. For example, substantial amounts of error in the test score of an individual may result if that person's performance is hindered by high levels of test anxiety or if he is not motivated to do his best or to respond truthfully to test items.

Anxiety, motivation and frustration. In the area of ability measurement, tests are typically geared to the ability level of the average member of a group. Such tests will be a rather different experience for examinees of differing ability levels. The low ability individual receives a series of items which are too difficult for him or her and may react by becoming threatened, anxious, or frustrated--the test may seem hopeless and he may simply stop trying. The high ability individual, on the other hand, receives items which are too easy for him--this person may find the task boring and unchallenging and, in a fashion similar to that of the low ability examinee, may simply stop trying to do his best. It is only for the average ability examinee that the items are likely to be sufficiently difficult to be challenging and yet not so difficult as to seem hopeless.

Adaptive testing procedures, however, tend to maintain an appropriate level of item difficulty for each individual. We don't yet know whether or not difficulty levels appropriate to each individual's ability level are the best ones for keeping motivation at high levels and anxiety and frustration at low levels. But at least adaptive testing procedures should keep the relative degree of item difficulty constant across ability levels and should thus have less tendency to arouse differential levels of motivation, anxiety, or frustration in individuals of different ability levels.

Feedback. Computerized test administration also makes it very easy to provide the examinee with feedback, immediately after each item response, as to the correctness or incorrectness of that response. A number of writers (e.g., Bayroff, 1964; Ferguson & Hsu, 1971; Zontine, Richards & Strang, 1972; Strang & Rust, 1973) have suggested that immediate knowledge of results, or feedback, may have positive motivating effects on some examinees and, therefore, may increase the likelihood that they will perform to their fullest capacities. Knowledge of results has long been considered important in the area of learning and instruction and has been built into methods of programmed and computer-assisted instruction. Further, the constructors of individually-administered intelligence tests, for example, Binet, Terman & Wechsler, all stressed that some form of encouragement by the examiner was essential in keeping the examinee motivated and performing to his fullest capacity, although this encouragement was *not* to include knowledge of results *per se*.

Since the effects of immediate feedback on performance on objective tests of ability have been only rarely studied, we have incorporated immediate feedback into some of our research designs.

Feedback and race.[2] In one study, both a conventional test and a pyramidal adaptive test were administered by computer to a group of inner-city high school students. The group was racially mixed, consisting of both black and white students. Tests were administered such that half the group received the conventional test first, while the other half received the pyramidal test first. Within each order of test presentation, half the group received feedback and the other half did not.

The results of the 3-way ANOVA for the conventional test scores are shown in Table 2, using number correct as the dependent variable. The only significant main effect was for race, with the overall performance of blacks being significantly lower than that of whites.

Table 2

3-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
ORDER	1	105.76	1.36	.25
RACE	1	2,013.26	25.84	.00*
FEEDBACK	1	81.74	1.05	.31
RACE X ORDER	1	161.54	2.07	.15
ORDER X FEEDBACK	1	28.74	.37	.55
RACE X FEEDBACK	1	170.40	2.19	.14
ORDER X RACE X FEEDBACK	1	599.40	7.69	.01*
ERROR	82	77.92		

However, the 3-way interaction among order, race, and feedback was highly significant. Figure 27 shows the means for the 3-way interaction. The left side of the graph shows the group means under feedback conditions, while the right side shows the means under no-feedback conditions. Note that the performance of whites was uniformly better than that of blacks *except* under feedback conditions when the conventional test was given first. In this case, the performance of blacks was not significantly different from that of whites.

Further analysis of this result suggested that it was due to motivational effects. If it can be replicated it suggests the possibility that under optimal conditions of test administration the performance differential between racial groups might be substantially reduced.

Feedback, ability level and testing strategy. In a second study, either a conventional test or a stradaptive test was administered with or without feedback in two groups of subjects. One group was a "high ability" group (College of Liberal Arts) and the other a relatively "low ability" group (General College) based on average college admission test scores and high school grades.

[2] These data were analyzed by Ms. Clara DeLeon.

Figure 27

MEAN SCORES FOR BLACKS AND WHITES COMPLETING THE 40-ITEM CONVENTIONAL TEST FIRST AND SECOND, BY FEEDBACK CONDITION

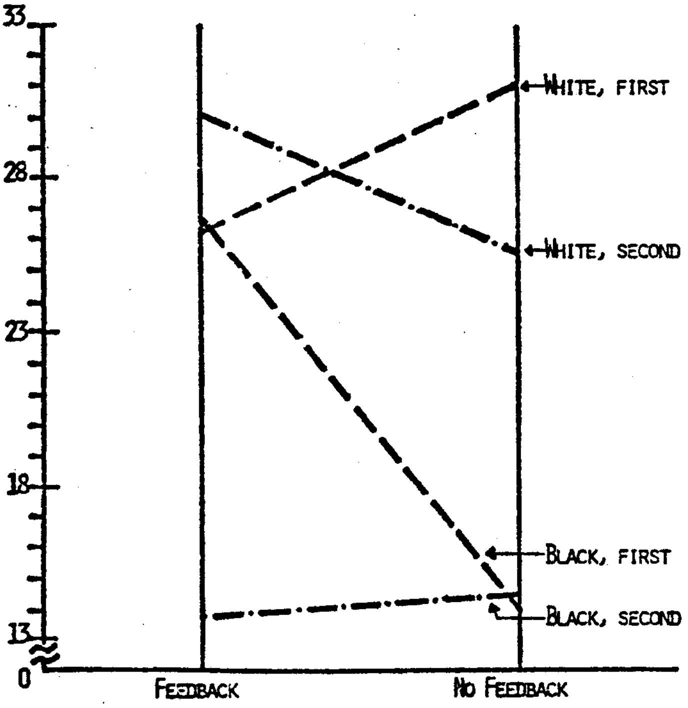


Table 3

MEAN NUMBER CORRECT ON 50-ITEM CONVENTIONAL TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

GROUP	FEEDBACK N	FEEDBACK MEAN	NO FEEDBACK N	NO FEEDBACK MEAN	TOTAL N	TOTAL MEAN
COLLEGE OF LIBERAL ARTS	60	30.47	57	27.10	117	28.83
GENERAL COLLEGE	28	22.54	28	20.71	56	21.62
TOTAL	88	27.94	85	25.00	173	26.50

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	1945.29	21.67	.001*
FEEDBACK	1	354.28	3.95	.046*
GROUP X FEEDBACK	1	22.45	.25	.999
ERROR	169	89.77		

Table 3 shows the mean number-correct scores on the conventional test according to whether feedback was or was not given. The analysis of variance indicated a significant main effect for feedback, and analysis of the means indicated that in both subject groups, the provision of feedback resulted in significantly higher test scores. For example, in the College of Liberal Arts group, the mean number correct under feedback conditions was over 30, while that under no-feedback conditions was only 27. A difference of 3.5 score points on a 50-item test could be highly influential in a practical decision about an individual.

The results for the conventional test showed that feedback had a positive effect on test performance, but when we looked at the stradaptive test, the results were quite different. Table 4 shows maximum likelihood scores on the stradaptive test under feedback and no feedback conditions. Note that there is no significant effect for feedback.

Table 4

ABILITY ESTIMATES FOR STRADAPTIVE TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.66	62	-.62	122	-.64
GENERAL COLLEGE	28	-.96	27	-.81	55	-.89
TOTAL	88	-.76	89	-.68	177	-.72

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	2.27	1.75	.184
FEEDBACK	1	.24	.19	.999
GROUP X FEEDBACK	1	.10	.07	.999
ERROR	173	1.29		

However, in trying to interpret these apparently conflicting results, it is necessary to remember that in the stradaptive test, almost everyone answers about half the number of items administered correctly; thus, the feedback should be about half negative and half positive. In the conventional test, however, high ability examinees receive mostly positive feedback while low ability examinees receive mostly negative feedback. Further, the stradaptive test maintains item difficulties at levels appropriate to each examinee's ability so it is perhaps a less stressful and more positive experience, particularly for "low ability" testees.

Further analysis revealed that the levels of motivation reported by examinees who took the stradaptive test were uniformly higher than the levels reported by those who took the conventional test. These data suggest that an adaptive test led to higher levels of motivation whether or not feedback was given. Thus, particularly for the low ability testees, an adaptive test may have the same motivational effects that giving feedback on a conventional test seems to have.

Implications. The results I have presented here are obviously not conclusive. Replications and further studies are certainly necessary. But given the current concern with test fairness and bias, it seems that we should pursue further the effects of various conditions of test administration upon performance. Adaptive testing and immediate knowledge of results may be able to provide testing conditions more conducive to allowing each individual to demonstrate his/her fullest capacities in test performance. And, since computerized adaptive trait measurement can provide us with important additional information of a variety of types, it has promise of supplementing the paper and pencil tests which have dominated psychological testing for the last 50 years.