# Student Attitudes Toward Tailored Testing

Bill R. Koch
Wayne M. Patience
University of Missouri--Columbia

As tailored testing procedures gain in popularity and are frequently applied in testing situations, it becomes increasingly important to determine the psychological aspects of the tailored testing environment which may be introducing error into the test scores (Weiss, 1975). On the surface, favorable attitudes toward tailored testing would be expected due to such inherent characteristics as self-pacing of progress through the test, reduction in test length (both time and number of items), and matching of item difficulties to individual ability levels. All of these features contrast with traditional multiple-choice tests.

Some concern has been voiced, however, regarding a possible increase in the examinee's anxiety level during a test that involves the novelty of interaction with the computer. In addition, a frequent complaint from students that test questions on traditional tests did not cover the material they knew (and hence did not measure their true abilities) might be amplified in tailored tests, where the total number of items administered may average only 10 rather than the 50 typical of traditional tests (English, Reckase, & Patience, 1977).

Unfortunately, little of the published tailored testing research has reported on measures of examinees' attitudes toward the procedures. One study (Hedl, O'Neil, & Hansen, 1973), which attempted to determine the attitudinal effects of computerized intelligence testing, employed a five-item scale to measure anxiety level. It also employed a brief attitude scale to measure preference for the computerized test as compared to examiner-administered tests. The results indicated significantly higher anxiety levels and significantly less favorable attitudes for the computerized test than for the regular examiner-administered test. However, these results were probably due to an artifact of the study, since the computerized test (non-tailored) was full length, and examinees were required to complete all items, even if they reached the test ceiling by failing 10 consecutive items. Examinees who had such a failure experience scored significantly higher in anxiety level than the persons who did not reach the ceiling on the test. Thus the overall findings of this study were distorted.

A subsequent study (Lushene, O'Neil, & Dunn, 1974) attempted to measure the congruent validity of the MMPI as administered by a computer compared to the traditional booklet form; an anxiety level measure was incorporated as well. The results of the anxiety data indicated that the computer test initially produced higher anxiety levels in the examinees than the booklet form, but that this anxiety quickly dissipated

once the computer session was underway; and there were no differences in anxiety levels at the end of the testing.

In this same area of personality assessment, there has been substantial research conducted using computers to administer personality instruments in an attempt to standardize the testing environment and eliminate the biases that may be induced with human examiners. Several studies hypothesized that examinees would respond more openly and honestly to highly personal or threatening items presented by the computer rather than by the human test administrator, but no significant findings have emerged in that direction (Resmovic, 1977). The proposed explanation for these results was that the studies failed to utilize the full interactive capabilities of the computer, using it instead as simply a presentation device.

Recently, Betz and Weiss (1976) conducted a comprehensive study which examined such attitude dimensions as level of motivation, anxiety level, perceived test difficulty, and immediate feedback of results on vocabulary tests. One important finding was that motivation was greater for low-ability examinees taking computerized adaptive (tailored) tests than for conventional tests administered on the computer. Another result was that significantly more anxiety was reported on the adaptive test than on the conventional test, even though both were computer administered. Also, students were able to perceive the difficulty of the test fairly well, although they were less consistently able to do so for the adaptive test. Finally, the immediate feedback feature was received very favorably by the examinees.

The attitude research on tailored testing has been conducted in relation to ability or personality testing rather than achievement testing. The difference, of course, is that in ability and personality tests, the examinees are typically asked to do their best or to respond honestly; but they have no clear incentive to do so. Achievement tests, on the other hand, are routinely used to assign course grades or for classification or placement decisions. One purpose of the present research, therefore, was to provide an indication of the attitudinal effects of tailored testing in the achievement test setting. Findings similar to previous research were expected in regard to perceived difficulty and computer interface effects; but differences were expected in such areas as anxiety and motivation levels, since in some cases achievement test results were used for course grades.

## Instrumentation

Two separate attitude questionnaires were administered during the course of the present studies. The first instrument was a Likert-type scale consisting of four statements which measured attitudes toward tailored tests on the dimensions of (1) time pressure, (2) perceived test difficulty, (3) test anxiety, and (4) general test preference. The questionnaire was administered subsequent to examinees' tailored testing sessions in three separate research studies. In each case the attitude measures were secondary to the overall thrust of the research.

The second instrument was a three-part attitude survey which was administered during the course of the fourth study. The initial section of the questionnaire consisted of four items. Each item asked the examinee to rank five different test modalities along the dimensions of (1) perceived difficulty, (2) time pressure, (3) anxiety or stress level, and (4) overall preference. For example, Item 1 read as follows:

1. Assume that you have a test coming up in some course that you are taking. For the 5 types of tests below, please rank them into an order of difficulty. The type of test that is most difficult for you should be ranked 5, while the easiest test should be ranked 1.
   _____ true-false test (paper-and-pencil)
   _____ essay test
   _____ multiple-choice test (paper-and-pencil)
   _____ oral examination
   _____ computer-administered multiple-choice test

The items for time pressure, anxiety, and overall preference were nearly identical in format to Item 1. The order of the five types of tests, however, was varied to reduce the likelihood of a fixed response set during the ranking procedure. The design of this section of the questionnaire was based on Coombs' (1964) unfolding theory: the items attempted to determine existence of a latent attribute underlying preferences for the five types of tests (an ordered metric scale) for each of the four attitude dimensions. The unfolding technique makes minimal assumptions in finding the order of the stimuli, as well as the size of the distances between them.

The second part of the questionnaire consisted of just three items in which the examinees' prior experience with computers was measured. The items read as follows:

Are you at all familiar with computers?
____ Yes _____ No
Have you ever punched computer cards at a keypunch machine before?
____ Yes _____ No
Have you ever interacted with a computer by means of a terminal before?
____ Yes _____ No

In the subsequent analysis of the data, the response scores to this section of the questionnaire were used as covariates with the four unfolding items in order to determine the effects of computer familiarity on the attitudes expressed.

The third part of the attitude survey consisted of a scale of six Likert-type items in which each statement was followed by the alternative responses from which the examinee was to choose. The purpose of this section of the survey was to determine the examinees's relative preference for a black-on-white compared to a white-on-black cathode-ray-terminal (CRT) display screen; however, the statements did not make any specific

reference to that purpose. This particular section of the survey was administered twice to each examinee: once after the tailored test in the black-on-white mode and once after the white-on-black test mode. Listed below are some examples of the statements:

1. The viewing screen was uncomfortable on my eyes.
   strongly agree agree neutral disagree strongly disagree
5. It was very easy to read the words and questions on the screen.
   strongly disagree disagree neutral agree strongly agree
6. Reading the questions on the screen was not much different from reading them on a regular test on paper.
   strongly agree agree neutral disagree strongly disagree

The response choices were weighted from 1 to 5 in the usual Likert fashion for scoring.

It should be noted that in neither of the two separate attitude questionnaires was the examinee responding anonymously, since the individual's student identification number was recorded at the time of the questionnaire administration.

## Tailored Testing Research Designs

The attitude data reported in this paper were collected as supplementary parts of four different experimental designs. The four-item Likert questionnaire was administered during the course of the first three studies, all of which compared computerized tailored testing to traditional paper-and-pencil achievement tests; the second attitude questionnaire was administered during the fourth study.

The initial study investigated the reliability and validity of tailored testing compared to traditional achievement tests (Reckase, 1977). The test content covered the statistics and measurement portion of an introductory course in educational measurement and evaluation at the University of Missouri. The study employed a test-retest design (test sessions one week apart) with the attitude questionnaire being administered after the second session. Although the tailored test did not count toward the students' grades in the course, the students did receive extra credit for their participation.

The second study (English et al., 1977) was primarily concerned with measuring differences in levels of achievement for students taking tailored tests compared to those taking traditional paper-and-pencil tests. The examinees were enrolled in an introductory course in educational measurement and evaluation at the University of Missouri. Again, a test-retest design (test sessions three weeks apart) was employed, with the attitude questionnaire being given to the examinees after their second session. In this case, however, the results of the tailored tests were used in the assignment of course grades, thus providing an evaluation of the procedure under motivated circumstances.

The third study investigated the effects on performance in achievement tests of paced versus self-paced scheduling of the time of tailored test administration. In addition, tailored test performance was compared with traditional paper-and-pencil tests. The self-paced groups could take the tests whenever they wished and as often as they liked until satisfied with their grades. In contrast, the paced and traditional groups were scheduled to take the test at a specific time and could take it only once. Again, the tests counted toward course grades, and the attitude questionnaire was administered subsequent to the second test session.

The final experiment was essentially a pilot study concerned with applying the one- and three-parameter logistic models to tailored achievement tests. The purpose was to check out programs and procedures in preparation for subsequent live-testing studies. A counterbalanced experimental design was employed in which each examinee had two test sessions approximately one week apart. If the examinee took the test for the first session on the black-on-white CRT, then the second session would be on the white-on-black CRT, and vice versa. Lighting conditions in the test room were held constant, as were the CRT screen brightness and contrast controls.

The three-part attitude questionnaire was first administered after the examinee's initial test session. The third part only (dealing with the CRT screen display mode preferences) was re-administered following the second session, yielding attitude data for each display screen format.

The achievement test itself dealt with the evaluation techniques section of an introductory measurement and evaluation course at the University of Missouri. All students had previously just completed the traditional paper-and-pencil test for this section of the course. Therefore, although extra credit was given for participation in the study, the tailored tests did not count toward course grades.

## Attitude Research Designs

In order to address the issue of examinees' attitudes toward tailored testing under unmotivated as compared to motivated conditions, a comparison was made between (1) the examinees' attitude responses to the questionnaire in the first study, where the tailored test did *not* count for course grades and (2) the responses on the same questionnaire in the second study, where the test did count. Obviously, this design was vulnerable to internal validity considerations, since the two groups were not based on random assignment, but were established depending on which semester the students took the course. The tailored test in both cases, however, covered identical course material and the physical testing conditions were equivalent (same test room and CRT terminal). Also, in both studies the examinees' participation was voluntary. This should have limited any possible differential selection effects of the first study compared to the second. Prior to analysis, the responses of an equal number of examinees from each group were randomly selected for comparison.

In addition to the analysis discussed above, each of the first three studies was subjected to a series of correlation analyses to measure the relationships between attitudes and such variables as ability levels, number of test items administered, and time spent taking the tailored test. Proportions of responses for each of the alternatives to the attitude items were also calculated to permit simple descriptive comparisons.

The second attitude questionnaire, which was administered as part of the fourth study, had three main research purposes. First, the responses to the Coombs' unfolding items were tabulated in order to determine the sets of individual preference rank orderings (called *I*-scales) of the five test types for each of the four attitude dimensions. These *I*-scales were then manipulated according to the unfolding technique to see if one dominant scale for each dimension (*J*-scale or joint continuum) could be recovered, upon which most of the respective *I*-scales would fit.

The second phase of the research was to convert the resulting *J*-scales from ordered metric scales into approximate interval scales, so that numerical values could be assigned to each of the positions along the *J*-scales. Upon completion, the scale values for each of the four attitude dimensions were related to ability levels, number of items administered, time spent taking the test, and prior computer experience, by means of multivariate analysis of variance procedures.

Finally, the last part of the questionnaire also used the multivariate analysis of variance technique to measure differences in responses to the six Likert-type items. The analysis compared attitudes toward the white-on-black CRT screen display to those toward the black-on-white CRT screen display. ADDS Consul 980 CRT terminals, which have both capabilities, were used in this part of the study.

## Results

### Motivated vs. Nonmotivated Groups

As can be seen in Table 1, the multivariate analysis of variance performed on the four attitude items for the motivated group compared to the unmotivated group yielded a statistically significant difference, approximate $F$ (4,69) = 6.249, $p < .001$. The cell means indicate that the differences between the groups were observed in relationship to the time pressure and anxiety dimensions. The subsequent one-way analyses of variance to compare the groups on these dimensions yielded $F$ (1,72) = 9.88, $p < .01$ for time pressure and $F$ (1,72) = 4.11, $p < .05$ for anxiety. No significant differences were found regarding perceived test difficulty or overall preference for the tailored test compared to traditional paper-and-pencil tests.

These results indicate that the examinees in the motivated setting (where the tailored test counted toward their course grades) felt that the tailored test had less time pressure than the unmotivated group did. This was accompanied by higher anxiety levels during the tailored test for the motivated group.

Table 1
Means and MANOVA for Attitudes in
Motivated vs. Unmotivated Settings

| | Cell Means | |
|---|---|---|
| Variable | Motivated | Unmotivated |
| Time Pressure | 2.73 | 2.22 |
| Difficulty | 1.97 | 1.76 |
| Preference | 1.95 | 1.78 |
| Anxiety | 1.68 | 2.00 |

Univariate Analyses of Variance

| Source | *SS* | *df* | *MS* | *F* |
|---|---|---|---|---|
| Time Pressure | | | | |
| Mot./Unmot. | 4.88 | 1 | 4.88 | 9.88** |
| Error | 35.57 | 72 | 0.49 | |
| Anxiety | | | | |
| Mot./Unmot. | 1.95 | 1 | 1.95 | 4.11* |
| Error | 34.11 | 72 | 0.47 | |

**$p < .01$;
*$p < .05$

Likert Attitude Items

The response data to the four attitude items have been summarized in Table 2 for the three studies in which the first questionnaire was administered. Again, it is interesting to note the differences in response percentages to the alternatives for each item for Study 1 (unmotivated) compared to Studies 2 and 3 (motivated). In addition, it is possible to observe the overall attitude responses of the examinees toward tailored tests as compared to traditional tests on the four items of the questionnaire.

For example, regarding the dimension of time pressure, the majority of unmotivated examinees felt equal time pressure for both types of tests, while the majority of motivated examinees felt less time pressure on the tailored test. In terms of perceived difficulty, 70% of the unmotivated examinees found the tailored test more difficult, while the motivated examinees tended to find the tests equally difficult. Opinion appears to be about equally divided for all the examinees regarding overall test preference, although there is a tendency toward preference for the tailored test. Finally, although the motivated examinees tended to find that the tailored testing aroused as much as or more anxiety than traditional tests, the opposite was true for unmotivated examinees.

The results of the correlational analyses were inconclusive; individual correlation coefficients varied substantially for a given pairing of variables across tests. Only a few consistent correlations emerged, such as the findings that high ability examinees tended to find the tailored test easy and that examinees receiving higher numbers of tailored test items tended to find the tailored test more difficult.

Table 2
Likert Attitude Items and Response Data

| Item | Response Percentages Study 1 ($N$=64) | Study 2* ($N$=41) | Study 3* ($N$=85) |
|---|---|---|---|
| 1. Compared to multiple choice tests, the tailored test has | | | |
| a. more time pressure | 19% | 7% | 11% |
| b. less time pressure | 26% | 78% | 66% |
| c. about equal time pressure | 55% | 15% | 23% |
| 2. Compared to traditional multiple choice tests, the tailored test is | | | |
| a. easier | 23% | 25% | 8% |
| b. harder | 70% | 31% | 27% |
| c. about as difficult | 7% | 44% | 65% |
| 3. As compared to the traditional multiple choice test, | | | |
| a. I would rather take the tailored test | 42% | 44% | 57% |
| b. I would rather take the traditional test | 25% | 44% | 29% |
| c. I prefer both equally well | 33% | 12% | 14% |
| 4. Taking the test on the computer makes me | | | |
| a. more anxious than a traditional test | 30% | 42% | 42% |
| b. less anxious than a traditional test | 45% | 12% | 21% |
| c. about equally as anxious as the traditional test | 25% | 46% | 37% |

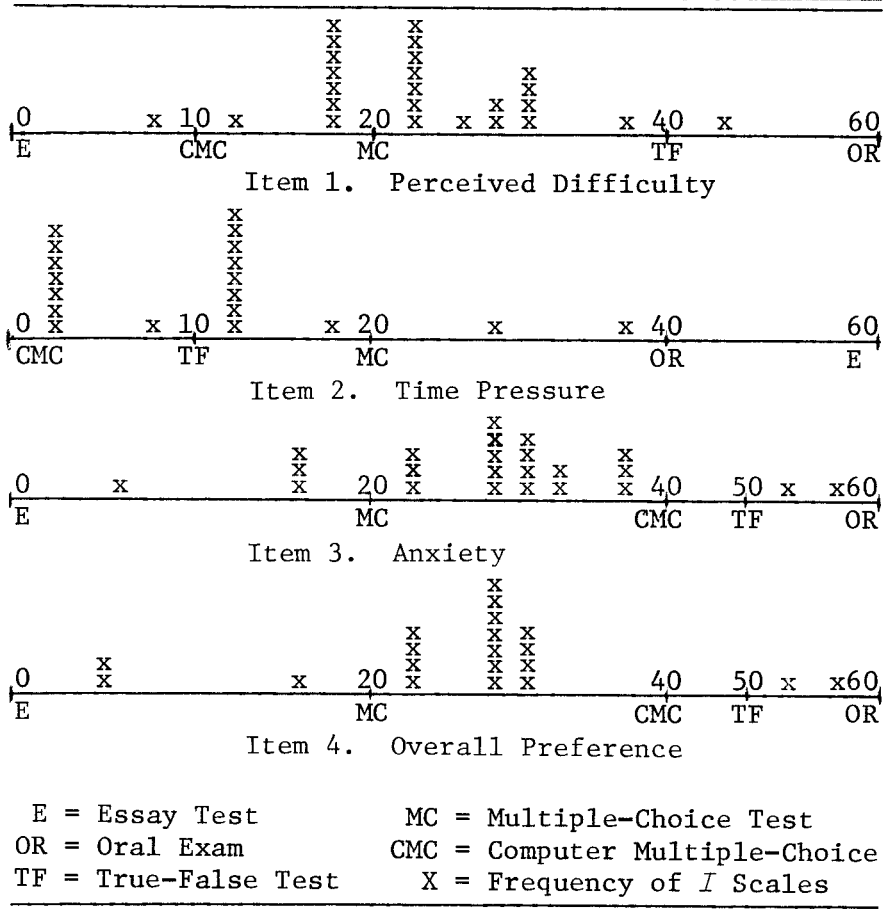
* In Studies 2 and 3 the tailored test counted toward the course grade, but not in Study 1.

Unfolding Items

The results of the analysis of the responses to the four Coombs unfolding items from the second attitude questionnaire are presented in Figure 1. First, it can be seen that only three dominant $J$-scales or joint continua were found for the four attitude dimensions, since the anxiety continuum is identical to the overall preference continuum. The $J$-scales have been converted from ordered metric scales into approximations of interval scales. This shows the order in which the five test types fall along each continuum, as well as the relative distances in terms of preference between the five tests on the scales (Coombs, 1964). In addition, the frequencies of the $I$-scales are indicated with "X's" above each $J$-scale. Each $I$-scale denotes the position of an examinee on the scale in regard to preference for each of the five tests; the closer a person's position to a test, the more it is preferred.

For example, there were two dominant $I$-scales for the Perceived Difficulty $J$-scale. A total of 14 examinees found multiple-choice tests (MC) to be least difficult and computer multiple-choice tests (CMC) to be slightly more difficult. Essay (E) and true-false (TF) were even more difficult, and oral exams (OR) were the most difficult.

Figure 1
*J*-Scales for Four Attitude Items




For the Time Pressure *J*-scale, there were again two dominant *I*-scales. One set of examinees found the computer multiple-choice test to have the least time pressure. In increasing order of time pressure, this was followed by TF, MC, OR, and E. The second group found TF to have the least time pressure, then MC, CMC, OR, and E.

The Anxiety *J*-scale showed the most variability in terms of test preference. No clearly dominant *I*-scales were evident, although there was a tendency for MC and CMC tests to have the lowest levels of anxiety for most of the examinees. On the Overall Preference *J*-scale, most of the examinees liked the MC tests best, the CMC tests second best, and the OR exams least.

Since score values were assigned to the examinees according to their positions on the four *J*-scales, a series of four separate multivariate analyses of variance were conducted between the *J*-scale score values and (1) low vs. high ability levels, (2) low vs. high number of items administered on the tailored test, (3) low vs. high amount of time spent taking the test, and (4) low vs. high levels of prior computer experience.

Overall significant findings resulted for only two of the analyses--number of items and amount of time spent taking the test, approximate $F$ (4,51) = 2.58, $p < .05$ and approximate $F$ (4,51) = 2.79, $p < .05$, respectively. However, none of the subsequent one-way analyses were significant in either case, making the findings difficult to interpret. It is clear that the attitudes of examinees taking many vs. few items or spending much vs. little time on the test differed significantly in terms of their scores on the four $J$-scale continua; however, where or how they differed is not clear. Perhaps part of the problem is related to the fact that none of the four $J$-scales was purely unidimensional. It was possible to fit only about half of the complete set of reported $I$-scales to any of the final four $J$-scales.

## Preferences for CRT Displays

The final multivariate analysis of variance was performed on the results from the six Likert items dealing with the examinees' preference for the black-on-white compared to the white-on-black CRT display screens used for the tailored tests. The results presented in Table 3 indicate that the examinees' attitudes toward the two display formats were significantly different, approximate $F$ (6,105) = 2.628, $p < .05$. The subsequent one-way analyses revealed that the differences occurred primarily on Items 1 and 5, $F$ (1,110) = 5.48, $p < .05$ and $F$ (1,110) = 8.34, $p < .01$, respectively. Both of these items refer specifically to reading difficulty experienced in taking the test. The examinees reported that the white lettering against a black screen background was significantly more uncomfortable on their eyes and was more difficult to read than the black-on-white screen format.

Table 3
Means and MANOVA for Attitudes

| | Cell Means | |
|---|---|---|
| Attitude Scale Item | Black | White |
| 1 | 3.125 | 3.679 |
| 2 | 3.750 | 3.821 |
| 3 | 3.382 | 3.436 |
| 4 | 3.945 | 3.982 |
| 5 | 3.564 | 4.036 |
| 6 | 3.055 | 3.109 |

Univariate Analyses of Variance

| Source | *SS* | *df* | *MS* | *F* |
|---|---|---|---|---|
| Item 1--"Hurt Eyes" | | | | |
| Black/White | 8.58 | 1 | 8.58 | 5.48* |
| Error | 172.34 | 110 | 1.57 | |
| Item 5--"Reading Difficulty" | | | | |
| Black/White | 8.04 | 1 | 8.04 | 8.34** |
| Error | 105.93 | 110 | 0.96 | |

**$p<.01$; $p<.05$.

## Discussion

One important, but not surprising, finding of this study was that the attitudes of examinees toward tailored tests were different in motivated test situations as compared to unmotivated test settings. If tailored achievement tests are to be used to classify or place individuals, to assign performance grades, or when a clear incentive for performance is present, the heightened anxiety levels of the examinees will be a factor. Of course, learning research suggests that heightened anxiety levels facilitate problem-solving performance for simple tasks, but inhibit performance on more complex tasks (Travers, 1972). Further research needs to be conducted in order to determine these effects in tailored achievement tests.

In this regard, it is interesting to note that the Coombs unfolding items yielded identical $J$-scales for the anxiety and overall test preference dimensions. This result may indicate that differences in anxiety levels for various types of tests dictated the examinees' overall preference levels for these tests. Thus anxiety may outweigh the effects of numerous other factors, such as time pressure or difficulty level, in terms of test preference.

In general, the tailored tests fared reasonably well compared to the other four test types measured on the $J$-scale attitude dimensions. In most cases the tailored tests were considered to be most similar to the traditional multiple-choice test format, which was the most preferred test type overall.

Finally, the CRT terminal with the white-on-black display screen was judged by the examinees to be significantly more difficult to read and harder on their eyes than the black-on-white display screen. Further research needs to be conducted in different settings and on different tailored testing tasks before it is possible to say whether or not reading difficulty actually interferes with test performance.

## References

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, 1976. (NTIS No. AD A027170)

Coombs, C. H. A theory of data. New York: John Wiley & Sons, Inc., 1964.

English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 1977, 9, 158-161.

Hedl, J. J., O'Neil, H. F., & Hansen, D. N. Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 1973, 40, 217-222.

Lushene, R. E., O'Neil, H. F., & Dunn, T. Equivalent validity of a completely computerized MMPI. Journal of Personality Assessment, 1974, 38, 353-361.

Reckase, M. D. Computerized achievement testing using the simple logistic model. Paper presented at the 1977 Annual Meeting of the American Educational Research Association, New York, NY, 1977.

Resmovic, V. The effects of computerized experimentation on response variance. Behavior Research Methods and Instrumentation, 1977, 9, 144-147.

Travers, R. M. W. Essentials of learning. New York: The Macmillan Company, 1972.

Weiss, D. J. Adaptive testing research at Minnesota--overview, recent results and future directions. Paper presented at the Conference on Computerized Adaptive Testing, Washington, DC, June, 1975.