THE QUADAFLEX TESTING FORMAT AS A PAPER AND PENCIL APPLICATION OF ADAPTIVE ABILITY MEASUREMENT THEORY

Nick Maruhnich/Florida State University

Psychometricians have known for a long time that the most effective test items are those which are appropriate for the ability level of the examinee. Conventional tests usually contain items which are selected so as to be appropriate for the average ability level of a group of students. This has caused less effective measurement for the examinee whose ability level deviates from the average. Adaptive ability measurement tests (e.g., tailored, pyramidal, flexilevel, stradaptive, etc.) have attempted to remedy this situation by providing items to examinees which are best suited for their ability level.

Adaptive ability measurement branches students to items of appropriate difficulty by using prior knowledge of the students' abilities and/or by using the information the examinees provide when responding to test items. For example, the high ability students are branched away from the easiest items, and the low ability students are branched away from the hardest items. Many different formats and strategies have been suggested and tried. All of the models have certain advantages and limitations, but none has had much empirical evidence of success in the paper and pencil mode of administration.

The most common variations among models are the number of stages, fixed or variable step sizes, and whether or not the model uses prior knowledge about the student. The following is a brief review of some of the advantages and limitations of the various models which helped to shape the quadaflex model that will be presented in this paper.

Two-stage models are the simplest of the adaptive ability testing strategies (Weiss, 1974, pp. 3-11). They generally have a first-stage routing test for the examinees which branches them to any one of two or more second-stage measurement tests (i.e., tests peaked near the ability estimate provided by the routing test). The routing test is intended to provide a rough estimate of the examinee's ability level, and each measurement test is intended to refine this approximation. The measurement tests are peaked at different difficulty levels and are designed to differentiate among the abilities of individuals within a narrower range of ability than the routing test.

One of the limitations of the two-stage testing model is that of the irreversible error which is incurred when an examinee is branched incorrectly from the routing test. Once an examinee is branched to the measurement test, s(he) remains there. For example, if an examinee is improperly placed in a measurement test too hard for him/her, s(he) cannot be branched downward to a measurement test with easier items. An advantage of the two-stage models is their adaptability to a paper and pencil mode of administration.

Multiple-stage models require many branching decisions and usually a computer for administration. These models can generally measure a wider range of abilities more effectively than conventional tests with considerably fewer items. But many of the models have the problem of recovery after a mistake in branching has taken place. This is most noticeable in the decreasing step size pyramidal models (Weiss, 1974, pp. 18-22). A mistake in the first branching decision cannot easily be corrected. For example, if a low ability student guesses correctly on the first stage of a six-stage Robbins-Munro shrinking step size pyramidal test (Lord, 1971a), then s(he) cannot be branched below items of median difficulty. Clearly, this is a serious flaw in the model.

The small, constant step size, multiple-stage models reduce the effects of guessing by not branching the examinees too far at any particular decision point (Weiss, 1974, pp. 12-17). This could also be regarded as a disadvantage when one considers the number of items it takes to branch a high ability examinee to the most difficult items. Variable step size models, with large initial steps, can remedy this by branching the high ability student to the more difficult items more quickly. Clearly, both fixed and variable step size strategies have limitations.

The only adaptive ability measurement that appears promising for paper and pencil test administration is the flexilevel format proposed by

Lord (1971b). This model starts the examinee at an item of median difficulty, then branches the examinee to an easier item for each wrong response and to a harder item for each correct response. The only empirical evidence (Olivier, 1974) available on the appropriateness of this model for paper and pencil administration has been negative.

Some adaptive testing models take advantage of prior knowledge of each student's ability. Weiss' (1973) stradaptive model is such a model. This model enters the examinee into one of nine ability strata by using either the prior information available on the testee or a self-report of ability (e.g., Weiss, 1973, p. 16). Other strategies that use prior information are the Bayesian and maximum likelihood formats, but they cannot be administered without the use of a computer (Weiss, 1974, pp. 56-67).

Quadaflex Tests

The quadaflex test is an attempt to integrate the paper and pencil mode of administration of conventional tests with adaptive ability measurement theory. The model borrows freely from many, but mostly from Lord's (1971b) flexilevel format and Weiss' (1973) stradaptive format. Basically, the flexilevel format has two strata, the quadaflex has four, and the stradaptive has nine. The branching rule for quadaflex is up-one stratum for a correct answer and down-one stratum for a wrong answer. If the student is in the basal or ceiling stratum, branching is to the next item in that stratum.

The optimum order of items within the strata is a matter for further investigation. The two forms that were tested in this study had alternating orders of item difficulties. For example, Quadaflex A had items in the 1st and 3rd strata arranged in an order of decreasing difficulty and the 2nd and 4th in an order of increasing difficulty. Quadaflex B had the opposite ordering of items (see Figure 1.)

Examinees on both forms start with the same

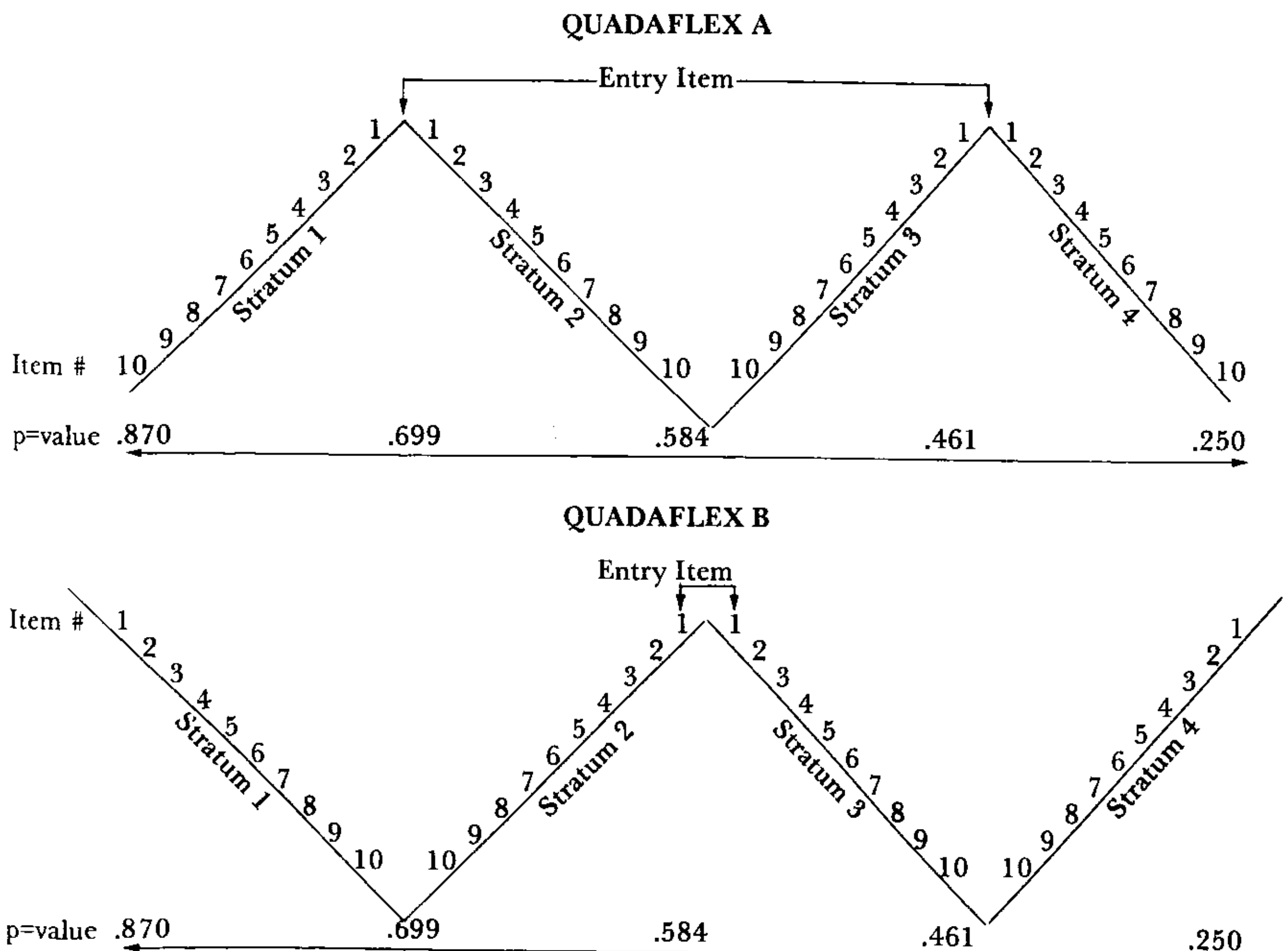

Figure 1. Quadaflex Forms A and B with different item orders within strata.

item. This is an item of median difficulty. Depending on the correctness or incorrectness of their answers, the students are branched appropriately.

Following are some examples:

(1) An examinee taking Quadaflex A and responding correctly to the entry item is branched to item 1 in stratum 3. This is the most difficult item in that stratum. If the examinee gets this item correct, the next item is item 1 in stratum 4. This is the next most difficult item on the test after item 1 in stratum 3. If item 1 in stratum 4 is correctly answered, the examinee is branched to item 2 in stratum 4. This process continues until the examinee misses an item in stratum 4 (or answers all the items in that stratum). For an incorrect response to an item, the branch is to item 2 in stratum 3. This is the lowest numbered item in that stratum that has not been attempted. Now, if the examinee misses item 2 in stratum 3, the branch is to item 1 in stratum 2. The examinee stops upon the completion of either 20 items or all the items in any one stratum.

(2) An examinee taking Quadaflex B and answering the entry item incorrectly is branched to item 1 in stratum 2. This is the hardest item in that stratum, and it is also slightly less difficult than the start item. If the examinee misses this item, the branch is to item 1 in stratum 1. This is the easiest item on the test. If the examinee answers this item correctly, the branch is to item 2 in stratum 2. If this item is answered incorrectly, item 2 in stratum 1 is the next item taken. If this item is missed, item 3 in stratum 1 is the next item. Again, the examinee stops upon completion of either 20 items or all the items in any one stratum.

It was conjectured that these patterns of item difficulties would be more effective for students of different abilities. Quadaflex A would be a more effective measure than Quadaflex B for students whose abilities were near the median or at the tails of the distribution. Quadaflex B would be a more effective measure than Quadaflex A for students whose abilities fell between the first two strata or between the last two strata. Response patterns for students whose ability levels fell in those areas of the distribution (except at the tails) designated to be more effectively measured by each particular form would follow a path of decreasing step sizes of item difficulties which converged near or at the students' ability levels. In other words, a student who was taking Quadaflex A and had an average ability level would probably miss most of the items in stratum 3 and answer correctly most of the items in stratum 2. This RWRW. . . response pattern would provide the student first with large step sizes and then with smaller step sizes, converging towards the student's ability level.

The complexity of most response patterns would make interpretation of the effects of different orders of item difficulties within strata extremely complicated.

Statement of the Problem

This study investigated: (a) the feasibility of the quadaflex testing format as a paper and pencil application of adaptive ability measurement testing theory, (b) the appropriateness of different scoring methods for quadaflex testing, (c) the effects of different item orders within each stratum on the distribution of scores, and (d) the relationships between the distribution of scores on the conventional test (Florida Eighth Grade Test—Vocabulary Subtest) and each of the quadaflex test forms.

Method

Subjects

Eighty eighth-grade students enrolled at the Developmental Research School (DRS) at Florida State University participated in the study. The school population is consciously selected to reflect, as closely as possible, the population of the county.

Item Formats

Quadaflex A and Quadaflex B each had 41 multiple-choice vocabulary items. Each item consisted of a short phrase containing an underlined word; below this was a column of four response alternatives. Each response alternative was one word. The examinee was to choose the alternative that was the best meaning for the underlined word. For example:

1. Postpone action
   A. assure
   B. simulate
   C. defer
   D. allot

Item Structure

Both forms of the quadaflex test had exactly the same 41 items. The median difficulty item from the pool of 41 items was the entry item for both forms. Each form also had four strata of items ranging from the 10 easiest items (stratum 1) to the 10 hardest items (stratum 4). The difference between the two forms was the order of item difficulties within each stratum. As discussed earlier, Quadaflex A had the items in strata 1 and 3 in an order of decreasing difficulty and the items in strata 2 and 4 in an order of increasing difficulty. Quadaflex B had just the opposite arrangement of items within each stratum.

All but one of the items were normed on the 1787 eighth-grade students in a large northern Florida county in February, 1975. The forty-first item was added in order to have 10 items per stratum. The item selected was one judged to be extremely hard for eighth-graders, and thus its p-value was estimated to be .25; this reflects the chance factor of one out of four examinees guessing the correct answer.

Test Materials

The test questions and answer sheets were color coded to help the examinee proceed smoothly from item to item. The entry item was purple, stratum 1 items were green, stratum 2 items were red, stratum 3 items were blue, and stratum 4 items were black. In boxes provided on the bottom of their answer sheets, examinees tallied the number of items completed as they proceeded through the test (see Figure 2). The test answer sheet was presented with the aid of a latent image and developing pen perfected by A. B. Dick Company. The examinees answered questions by marking the appropriate boxes with their developing pens. The pen reacted with the latent image to expose directions for choosing the next item to be taken in the sequence. The directions were simply the name of the color of the problem that the examinee was to respond to next. If the examinee answered this subsequent item correctly, the branch was up a color to the lowest number problem not previously attempted in that color. If the examinee answered the item incorrectly, the branch was down a color to the lowest numbered problem not attempted in that color. Finally, if the students were in the ceiling or basal stratum, they were branched to the next item in that stratum.

Mode of Administration

The test was administered on 2 consecutive days to 3 classes of eighth-graders. Each test administration took one class period. Before the testing began, students were randomly assigned to take either Quadaflex A or Quadaflex B and then given a quick lecture on adaptive ability

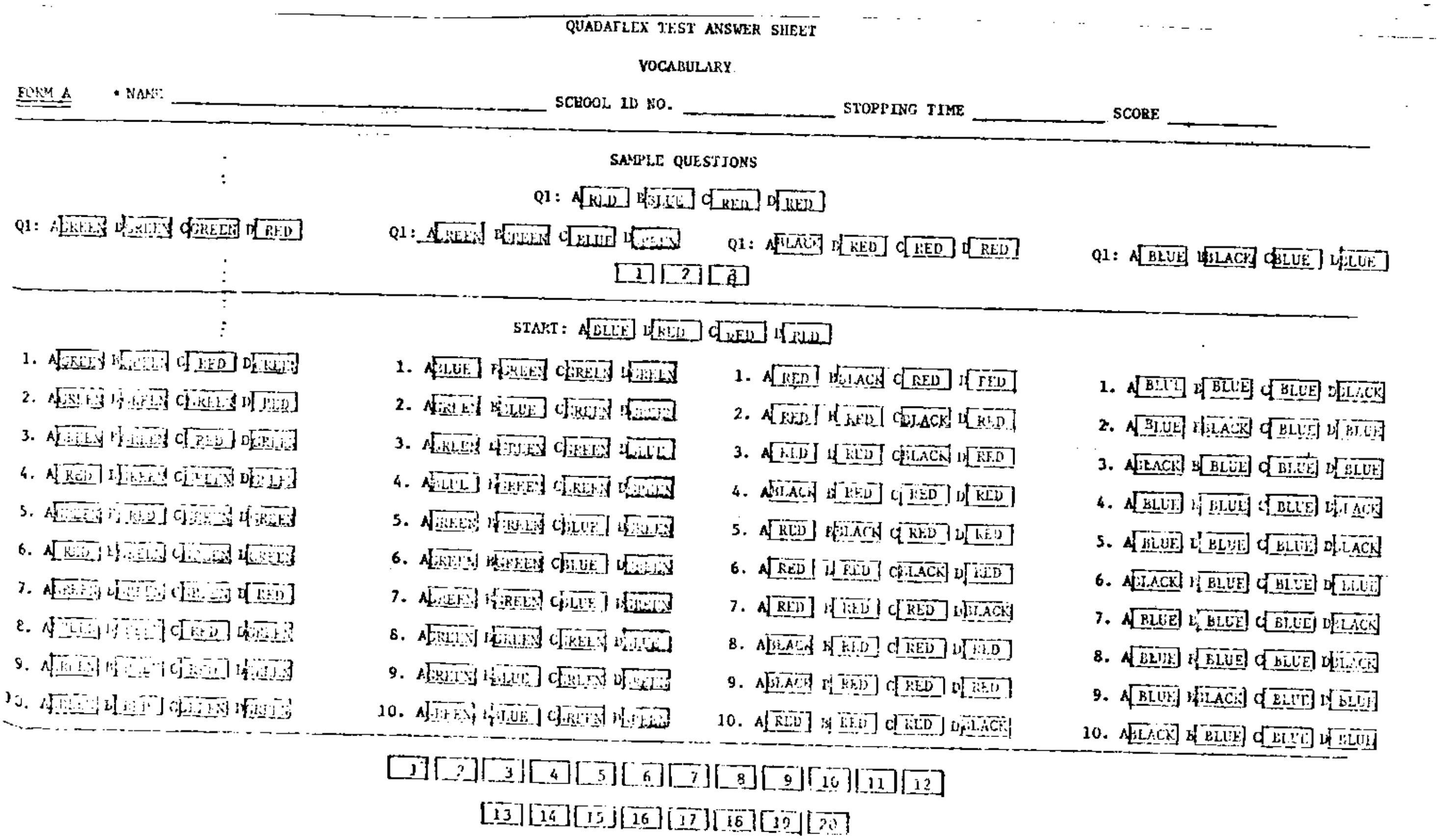

Figure 2. Example of the quadaflex test answer sheet.

measurement theory in terms they could understand.

Four graduate students, including the developer, served as proctors for each administration. The reason for the high proctor/examinee ratio was to help insure a minimum of mistakes which might be caused by the examinee's being unable to read his exposed latent image on the answer sheet. This added precaution seemed to have helped to keep the number of invalid tests to 2 out of 80. Olivier (1974) reported that his administration of a flexilevel vocabulary test to eighth-grade students yielded a ratio of invalid tests to valid tests of 76:655 or approximately 12%. Because of the relatively high ratio of invalid to valid tests that he encountered, he recommends that examinees practice responding to items in the flexilevel format before the administration of the actual test in order to familiarize themselves with its novel mechanics. He contends that this exercise will prepare them so that they will not be distracted by the unconventionality of the method and will attend wholly to the task of responding. However, the use of practice tests was deemed infeasible in this instance because the required time to drill the examinees is considerable. The high proctor/examinee ratio was proposed as a viable alternative and seemed to be an effective course of action for reducing the number of invalid tests.

When the students finished, they were instructed to raise a hand and call a proctor. The proctors marked the finishing time, circled the student's last answer, and collected the materials.

Scoring

A number of methods for scoring adaptive ability measures have been published in the literature. The appropriateness of the different strategies is a matter for further inspection. The scoring methods discussed here have been used with the flexilevel test, the stradaptive test, and the pyramidal test. They provided essential information in the search for a method suitable for scoring the quadaflex test.

In spite of the fact that the advantages and limitations of the various scoring methods will be pointed out with respect to particular tests in the following paragraphs, it is important to understand that these same advantages and limitations can and do in some cases apply to other tests. The characteristics of the scoring methods and the particular tests that they are associated with are provided to assist in illustrating those relationships and not to imply that they are unique to the example offered.

Lord (1971b) scores the flexilevel test by counting the number of items answered correctly; examinees who miss the last item, however, receive a one-half point bonus which adjusts for the (n+1)*th* item on an n-item test. For instance, if two examinees have equal number right scores but one misses the n*th* item while the other gets it right, the second examinee must get a harder item right on his next attempt in order for both to again achieve identical number right scores. Therefore, the second examinee's performance is, according to Lord, inferior to that of the first examinee. Unless an adjustment in scores is made, both examinees will obtain the same total score. The application of this bonus results in the first examinee's receiving a higher score. Lord (1971b, pp. 150-1) offers a complete justification for the use of this scoring method. One advantage of employing his technique is that it utilizes the information gained from the last item.

Unfortunately, however, Lord's scoring method fails to differentiate among examinees who obtain equal number right scores (adjusted for the last item), but who do not respond correctly (or incorrectly) to precisely the same items. For example, one examinee may answer the first item right and the remaining ones wrong; the second examinee may answer the first item wrong, the second item right, and the remaining ones wrong. Both examinees are classified by Lord (1971b, p. 150) as "red" examinees with equal number right scores. Since both examinees respond identically to all but two items, this scoring method seems inequitable because on those two items the second examinee responds correctly to the easier one and incorrectly to the more difficult one, whereas the first examinee responds incorrectly to the easier one and correctly to the more difficult one.

Weiss (1974, pp. 50-52) has suggested various methods for assigning scores to examinees, several of which are listed below. In some of these he employs the expression *stratum difficulty*, which he identifies as the average difficulty of all items in a particular stratum.

One's score is defined as (a) the p-value of the most difficult item answered correctly, (b) the p-value of the item that the examinee would have answered next if testing had continued, (c) the p-value of the most difficult item answered correctly in the highest non-chance stratum, (d) the average p-value of all items in the most difficult stratum in which an item was answered correctly, (e) the stratum difficulty at which the (n+1)*th* item would have occurred, (f) the average p-value of all items in the stratum immediately below the ceiling stratum, (g) the interpolated distance between his highest non-chance stratum and the next higher or lower stratum defined with respect to p-values and proportion correct in that stratum, (h) the average p-value of all items answered correctly, (i) the average p-value of all the items answered correctly between the basal and ceiling strata, and (j) the average p-value of the items answered correctly in the highest non-chance stratum.

A deficiency which is common to all of these scoring methods is their failure to consider the path the examinee follows through the test. For example, an examinee who guesses correctly the median difficulty or entry item and then answers incorrectly all remaining items on the test receives a score of average when evaluated by the method identified as letter h above. Clearly, this is not an accurate reflection of that examinee's ability.

Most pyramidal models use variations of scoring schemes which are based on item difficulty values. However, the techniques which define a score as a p-value of the final item attempted or the average p-value of all items attempted do not distinguish between the examinee who responds correctly to the final item and the one who responds incorrectly to it. As indicated previously, this is a weakness which plagues most methods developed thus far for scoring adaptive ability tests. Of those investigated here, the two most acceptable definitions of a score on an adaptive ability measure are: (a) the p-value of the (n+1)*th* item (method b above) and (b) the average p-value of the n+1 items attempted. Both of these methods differentiate between the examinee who answers the final item correctly and the one who does not.

The method to be employed for scoring the quadaflex testing format is the average difficulty of the n+1 items attempted on the test. This method recognizes the importance of the particular path an examinee pursues through the items and also uses the information generated from the last item. Because a low p-value is associated with a hard item and a high p-value with an easy one, the lower one's average p-value or final score, the more one presumably possesses of the attribute being measured.

Results

The descriptive statistics reported in Table 1 were computed for scores produced by averaging the difficulty values of the n+1 items attempted (see Table 1).

TABLE 1

Descriptive Statistics for Quadaflex Tests

	Quadaflex A	Quadaflex B
Mean	.495	.464
Variance	.006	.004
Minimum	.387	.394
Median	.466	.439
Kurtosis	-.431	.499
Maximum	.684	.638
Standard Dev.	.079	.060
Skewness	.745	1.155
Sample	37.	41.

The mean score or average p-value on Quadaflex A is significantly ($p < .01$) higher than the mean score on Quadaflex B. Perhaps these dissimilar mean scores are an indication that one or more of the following hypotheses are true: (a) Quadaflex A is biased downward, (b) Quadaflex B is biased upward, and (c) examinees taking Quadaflex B are superior to examinees taking Quadaflex A.

A test which discriminates more efficiently among examinees of average ability should render observed scores which cluster above and below that point. This should result in a platykurtic distribution which is precisely what happened: the statistics demonstrate that the distribution of scores for Quadaflex A is platykurtic (see Table 1).

A test which measures examinees between the first two strata and the last two strata more effectively, as Quadaflex B was hypothesized to do, would force scores to collect near the middle. This would produce a leptokurtic distribution instead of a normal one. Again this type of distribution was observed (see Table 1).

The statistics indicate a much larger positive skewness in the distribution of scores for Quadaflex B than for Quadaflex A and are consistent with the original hypotheses. In other words, Quadaflex B is hypothesized to be more efficient for students whose abilities fall between the last two strata and therefore would force a positively skewed distribution of scores with a mean near that point (between the last two strata) even further in the positive direction. This is exactly the type of distribution that was observed (see Table 1).

Interpreting the statistics in Table 1 is complicated for several reasons, including: (a) each distribution has a different median value, (b) the means for each distribution are dissimilar, and (c) the sample may be characteristically different from the norming population.

The Spearman rank correlation coefficient was computed for students' ranks on the conventional test and their ranks on either of the two quadaflex forms; both correlation coefficients were .85.

The most time required to complete either of the quadaflex tests was 12 minutes, whereas the time required to complete the conventional test was 20 minutes.

Conclusions

The high correlation between the students' ranks on the conventional vocabulary test and each of the quadaflex test forms seems to indicate that both types of exams are measuring examinees in the same way. This, coupled with the fact that both quadaflex forms provided a savings in testing time of nearly fifty percent, makes the quadaflex testing format an attractive alternative to conventional testing. In addition, the quadaflex testing format has demonstrated a lower rate of test invalidity (i.e., tests that had to be thrown out) than any other reported attempts to apply adaptive ability measurement theory to the paper and pencil mode of test administration. Certainly, the quadaflex testing format is an adaptive ability measurement technique which deserves consideration as an alternative to both conventional tests and other forms of adaptive ability tests.

References

Lord, F. M. Robbins Munro procedures for tailored testing. *Educational and Psychological Measurement,* 1971, *31,* 3-31. (a)

Lord, F. M. The self scoring flexilevel test. *Journal of Educational Measurement,* 1971, *8,* 1477-151. (b)

Olivier, P. *An evaluation of the self-scoring flexilevel tailored testing model.* Unpublished doctoral dissertation, Florida State University, 1974.

Weiss, D. J. The stratified adaptive computerized ability test. *Research Report 73-3,* Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973. (AD 768376)

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