DOCUMENT RESUME

ED 307 313 TM 013 441

AUTHOR	Gershon, Richard C.
TITLE	Test Anxiety and Item Order: New Parameters for Item Response Theory.
PUB DATE	Mar 89
NOTE	29p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Ability; Adaptive Testing; Adolescents; Adults; *Aptitude Tests; Comparative Analysis; *Difficulty Level; *Latent Trait Theory; *Personality Measures; Personality Traits; Predictive Validity; *Test Anxiety
IDENTIFIERS	*Item Position (Tests); Johnson O Connor Aptitude Tests; Rasch Model

ABSTRACT

Examinees (N=1,233) at the Johnson O'Connor Research Foundation (JOCRF) were administered one of three test forms in which only item order differed. The study was undertaken to determine the validity of the assumption underlying item response theory (IRT) that there are fixed item parameters that can predict performance. The Rasch IRT model was chosen. The three experimental tests were constructed from 950 items found in the JOCRF's item bank. The population seen at the JOCRF ranges in age from 14 to 60 years. Personality tests administered to subjects included Mandler and Sarason's Test Anxiety Scale and a short series of test-taking strategy items. All subjects took a minimum of 18 aptitude tests. Three primary factors were included in the analyses: (1) level of test anxiety; (2) ability; and (3) difficulty order. Results of the study indicate that item difficulty order, test anxiety, and ability all affect performance, violating assumptions of IRT. It appears that the theory neglects to address the effects of individual differences on test behavior. Adaptive testing techniques are reviewed, and the theories of test anxiety and associated measurement instruments are discussed. New testing strategies are proposed in which personality variables and test characteristics can be incorporated as parameters, allowing tests to be manipulated in a manner that maximizes performance. Three tables and nine graphs present study data. A 55-item list of references is provided. (TJH)

Test Anxiety and Item Order
New Parameters for Item Response Theory

Richard C. Gershon

Johnson O'Connor Research Foundation

and

Northwestern University

Paper Presented at the Annual Meeting of
The American Educational Research Association

March 30, 1989

Abstract

Item Response Theory (IRT) assumes the existence of fixed item parameters which can predict performance. The theory neglects to address the effects of individual differences in test behavior. Examinees at the Johnson O'Connor Research Foundation were administered one of three test forms in which only item order differed. Item difficulty order, test anxiety and ability were all found to affect performance, violating the assumptions of IRT. New testing strategies are proposed in which personality variables and test characteristics can be incorporated as parameters, allowing tests to be manipulated in a way which would maximize performance.

Test Anxiety and Item Order
New Parameters for Item Response Theory

> An examinee's response pattern on a test can be viewed as a product of a three-way interaction between the characteristics of the individual, of the test, and of the situation in which the test is given. (Birnbaum, 1986)

Ability testing has become an increasingly important feature in the lives of children attending school, adolescents applying to colleges, and adults seeking employment. Modern test construction uses Item Response Theory (IRT) to decrease test length and increase overall reliability in all of these milieus. The theory is used to compute item parameters to predict likelihood of a person answering a given item correctly. The model presumes fixed parameters, and that the likelihood of the person answering the item correctly is based only on ability. However, contemporary psychometricians have not addressed the impact of individual differences other than ability on test behavior, a mistake earlier discovered in implementing conventional test theory. Personality researchers have demonstrated that test performance is affected by test anxiety as well as changes in item order. If this is true, a major assumption of IRT is incorrect, and the value of a limited set of fixed parameters placed in question.

This study will review contemporary testing literature and introduce the fundamental concepts of item response theory. An overview of the adaptive testing literature will also be included as a premise for understanding how alternative test strategies can be affected by personality factors. The field of test anxiety, from its initial conception as a part of general anxiety theory, to its current research directed toward understanding how various levels of anxiety affect test performance will then be presented. Finally, a strategy will be proposed and implemented to examine the effects of anxiety, item difficulty order and ability on performance.

Test Theory

Authors of conventional tests are continuously confronted with the problem of "bandwidth versus fidelity" (McBride, 1976). For every test which they construct, authors must resolve the dilemma of where one places the peak efficiency of the test. Most tests are designed to be applicable to either the largest segment of the population (e.g., the population mean), or a select targeted population based upon a particular level of a given trait. In the former instance, a significant number of items will have to be concentrated at the population mean in order to preserve the reliability of the test as a whole. This will result in low reliability for subjects deviating significantly from the population mean.

A better approach to conventional test construction would be to maximize reliability at every point along the distribution. Theoretically this would entail concentrating a significant number of items at each possible value of the trait being tested. But this leads to problems with testing inefficiency. Given a 500-item standardized achievement test, the average person will have to attempt 250 items which are beyond his or her ability level. Conversely, high-ability subjects will have spent the bulk of their time filling in items which are far below their ability level.

Adaptive Testing Techniques

Adaptive testing refers to a technique which attempts to pinpoint the subject's functioning level by administering items as close as possible to the subject's true ability level, even though the level is not known in advance. The true functioning level is arrived at through a process of systematically narrowing down the possible range of abilities. This process results in tests with an average of seventy percent fewer items than conventional tests.

Adaptive testing can be conceptualized as a form of binomial search; successively administering items which narrow down the individual's true range of ability (or trait level). Given a large item bank, with good discrimination values, items can be administered until a satisfactory level of reliability is attained. Near perfect reliability is possible at any level of performance. Conversely, when only rough measurements are needed, only a small number of items need be administered.

Item response theory is used to compute the item parameters necessary for computerized adaptive testing (CAT) techniques. The advent of microcomputers allows inter-item calculations to be performed at speeds which allow for more efficient and more reliable test administrations. While similar calculating abilities were previously available on large mainframe systems, the associated costs made the use of adaptive techniques prohibitive.

CAT procedures require the storage of large banks of items and their associated parameters. Items with maximum information values are necessary for adaptive testing to be truly effective. Fortunately, latent theory can be used to convert existing "conventional" item banks for use in IRT applications (Urry, 1977). Computer packages are available which use the data commonly found in existing item banks to compute the IRT parameters (Hambleton & Rovinelli; Urry, 1977; Wood, Wingersky, & Lord, 1976).

While the advantages of computerized testing techniques may be fairly obvious, the recent appearance on the testing scene of these techniques has not yet allowed for sufficient research into their possible disadvantages and weaknesses. Many of the problems which occur in conventional testing situations may be magnified by the relative shortness of adaptive procedures. For example, item context and position have been found to violate the assumption of unidimensionality which is necessary for IRT applications. This may be a problem for tests which are susceptible to location or practice effects such as the analytic subtest of the Graduate Record Exam (Kingston & Dorans, 1984). Other differences might be found to occur as a function of the time differences found between different testing procedures (Moreno, et al. 1984). In addition, the absence of "warm up" items may also have an effect.

Adaptive testing can be approximated by the use of tailored tests administered without the use of a computer. Three types of tailoring are commonly used: status tailoring, performance tailoring, and self-tailoring (Wright & Stone, 1979). In each case the goal is to narrow down the range of items which the examinee must take. First, an approximate level of performance is determined, and then a test form is administered which examines only a small band of performance. Status tailoring is used when grade or age information is sufficient to determine approximate ability, while performance tailoring administers a pretest to estimate performance. In both cases, the difficulty of each form has been predetermined, allowing the difficulty of the main test to match the estimated ability range of the subject. Self-tailoring allows subjects access to the entire item bank, instructing them to begin with the first item which appears to be at their level, and to continue responding until the items seem to be beyond their ability level. All three tailoring methods result in shortened overall test length.

A good item bank is necessary for any tailoring strategy to be successful. Items must be gathered which evenly "blanket" the difficulty range of the entire battery. While the three-parameter model described earlier is a good method for computing item statistics, one- and two-parameter models have also been used successfully. The one- and two-parameter models have strict assumptions regarding guessing, and the one-parameter model also has assumptions regarding item discrimination. Items which do not match these criteria are usually eliminated from the bank.

While there are major theoretical differences between the application of conventional test theory and item response theory, both models should theoretically produce similar estimates of ability for the same person. However, proponents of IRT have yet to address the affects of individual differences, other than ability, on performance. Anxiety serves as an example of a personality variable which has been shown to affect performance. The next section of this paper will review theories of trait anxiety, discuss the underlying mechanisms which cause test anxiety, and examine how varying levels of test anxiety differentially affect performance.

Test Anxiety

In his review of anxiety and how it relates to testing situations, Matarazzo (1967) distinguishes "trait-anxiety" from "state-anxiety" as follows: trait anxiety refers to a chronic anxious condition in which the person is overwhelmed by anxious tendencies on a continuous basis, whereas state anxiety is situation dependent. One particularly anxiety-provoking situation arises when many individuals are faced with a test taking situation.

Test anxiety theory was first presented by Mandler and Sarason as the mechanism underlying their Test Anxiety Questionnaire (TAQ). Previous measures of anxiety had looked only at general measures of anxiety, while the TAQ was implemented to look at a singular anxiety state (Wine, 1971). Sarason defined test anxiety, "as a form of self-preoccupation--characterized by self-awareness, self-doubt, and self-depreciation--that influences overt behavior and psychological reactivity" (Sarason and Stoops, 1978, p. 103). For Sarason, performance decrements due to test anxiety are caused by a situationally induced cognitive style which interferes with normal performance.

Highly anxious individuals also have symptoms of decreased self-esteem, manifested in decreased expectancies for success (Doctor & Altman, 1959). In addition, more anxious subjects have also been found to attribute more blame to themselves in failure situations than do less anxious subjects (Doris & Sarason, 1955).

Studies have shown overall academic performance, as measured by "classroom tests, grade point averages, intelligence and aptitude tests" is less for more test anxious students than for less test anxious students (Deffenbacher, 1978, p. 248; Hollandsworth, Glazeski, Kirkland, Jones & VanNorman, 1979; Mitchell & Ingham, 1970; and Spielberger & Katzenmeyer, 1959). More test anxious students also have poorer study habits (Anastasi, 1982).

Testing and test anxiety are universal concepts in our society. Our culture is extremely test conscious. Test performance can often be seen as having life long effects, and therefore it is reasonable to expect everyone to exhibit some form of test anxiety (Sarason & Palola, 1960). In this regard, moderate test anxiety is a normal, necessary function of personality.

Theories of Test Anxiety

Attentional theory conceptualizes test anxiety in terms of types of responses. While low-anxious (LA) individuals elicit "task completion responses" the highly anxious evoke responses designed to avoid dealing with the aversion stimulus (i.e. an examination). The avoidance response either produces thoughts of inevitable failure (Child, 1954; Nicholls, 1984; Sarason & Mandler, 1952), or it changes the individual's focus of attention (Wine, 1971). High-anxious (HA) persons react to the aversive stimulus by attending to irrelevant task elements (Mann, 1972; Meichenbaum, 1972). Low-anxious individuals take advantage of their own level of anxiety and are driven to perform. Less anxious subjects will always outperform more anxious subjects due to the inability of more anxious subjects to allocate sufficient concentration to the task at hand.

While in every case, the individual is attending to in. [illegible] ioughts of varying worth to the testing situation, HA and LA subjects differ greatly in the gener[illegible] to which they are attending (Wine, 1971). Highly anxious individuals are frequently thinking of ways to escape from the testing situation (Galassi, Frierson & Sharer, 1981). Feelings are of "inadequacy, helplessness, heightened somatic reaction, anticipation of punishment, and loss of status and esteem". In essence, the thoughts are "self rather than task centered" (Mandler & Sarason, 1952, p. 166). In contrast, low test anxious individuals switch their attention/drive to performing the task at hand.

Cue Utilization Theory (Easterbrook, 1959) sees anxiety as a component of emotional arousal which affects performance. Within this model, highly anxious subjects should outperform their less anxious counterparts on tasks entailing few cues. In this case the low-anxious subjects are thought to over-incorporate irrelevant information, and success is limited. This performance differential is reversed when the tasks contain many cues, with LA subjects achieving superior performance, and HA subjects showing worsened performance as they exclude necessary cues.

Working Memory Capacity Theory (Eysenck, 1979) predicts anxious individuals differ in the amount of memory which they are able to devote to relevant information processing. In situations requiring memory use for task relevant information, this theory predicts that LA individuals will outperform their HA counterparts.

Leon & Revelle (1985) compared the utility of Attentional Theory, Cue Utilization Theory, and Working Memory Capacity Theory in terms of their abilities to explain the relationship between anxiety and performance. They found Attentional Theory to be the only anxiety-performance theory to be supported, at least when ego-involving threats were used as a stressor. No support was found for Working Memory Capacity Theory, and the support found for Cue Utilization Theory was attributed to a scaling artifact. None of the models were found to be conclusive when varying levels of trait anxiety were considered.

Caution Mechanism. Ruebush (1960) conceptualized test anxiety in terms of a caution mechanism. Caution has become an overlearned defense mechanism which is triggered by testing situations. Performance efficacy will be determined relative to the value of using caution in the given performance task. In a stuay designed to evaluate this theory, HA subjects were found to have much higher cautiousness scores. When these same subjects completed a task in which caution was an asset, the HA group outperformed the LA group. However, a large degree of caution is rarely necessary in testing situations, LA subjects in general, outperform their HA counterparts.

Worry and Emotionality -- Underlying Factors. Worry and emotionality are two important components in the test anxiety literature which have not been fully integrated into any of the popular theories of test anxiety per se. Wine (1971) summarizes the two terms by describing the worry component as a cognitive concern for performance, and emotionality as the autonomic arousal component in anxiety. These two constructs are viewed as being critical factors in all research which relates anxiety to performance (Morris & Liebert, 1970).

Three reviews have found distinctions between worry and emotionality in the test anxiety literature. Worry was found to be negatively related to performance and performance expectations (Deffenbacher, 1977; Deffenbacher, 1978; Meichenbaum, 1972). Worry appears to fit in best with an attentional interpretation of test anxiety. Worry is the component which directs attention away from the task at hand inwards to thinking about oneself (Wine, 1971).

Necessary Aspects of Test Anxiety

Some researchers have argued for the necessity of at least a small level of test anxiety (Mandler & Sarason, 1952; Ruebush, 1960). According to these researchers, performance can occur only when anxiety facilitates task relevant responding (Wine, 1971). In addition, some degree of anxiety increases mental alertness in many individuals, and for some learning is improved when moderate levels of test anxiety are continuously present. There are also testing situations where good performance is dependent upon a high level of test anxiety (Kirkland, 1971).

The quantity of test anxiety appears as an issue to parallel arousal theory. Yerkes and Dodson (1908) found support for an inverted-U theory of stimulus-performance. Their findings suggested there is an optimum level of arousal for any given task. Too much, or too little, arousal decreases performance. In the case of test anxiety, optimal drive level appears to be relatively low, except when caution is an important component. The drive level for LA subjects appears to remain constant throughout an examination, it is only when LA individuals believe the importance of the task is diminished that drive level decreases to the point that performance diminishes (Paul & Eriksen, 1964).

Measurement

The first general scale to measure manifest anxiety was introduced by Taylor in the early fifties. The purpose of this first test (The Taylor Manifest Anxiety Scale) was to measure Hullian drive strength and how it directly or indirectly influenced performance (Child, 1964 and Sarason, 1960). Shortly thereafter, Mandler & Sarason introduced the Test Anxiety Questionnaire. Their goal was to present a test which would measure the likelihood of becoming anxious in testing situations. Sarason later constructed a similar True-False measure which he entitled the Test Anxiety Scale (TAS) (Wine, 1971).

Research has found that while the TAS correlates with verbal aptitude, the Taylor Manifest Anxiety Scale does not (Alpert & Haber, 1960). In fact, numerous studies have found the two tests to measure different things (Sarason, 1960). In the case of the TAS, a recent analysis of the existing literature has determined the "active ingredient" is perceived competence (Nicholls, 1984).

Differential Effects of Test Anxiety

Test anxiety has differential effects dependent upon the level of anxiety and varying situational elements commonly found in testing situations. These conditions can be loosely grouped into 1) environmental factors and 2) test characteristics. Environmental characteristics are those elements of testing situations which are dependent upon the testing situation. These factors include externally introduced stresses (evaluation emphasis and time pressures), the presence of observers, the presentation of feedback (Anastasi, 1982; Meunier & Rule, 1967; Sarason, 1958), or the level of aspiration (Mandler & Sarason, 1952; Trapp & Kausler, 1958). Test characteristics refer to those elements of the testing situation which are found within the testing protocol itself including: test difficulty, the order of test items, the associative values of those items, the effect of repeated trials within a given testing situation, and the effects of tests which contain initial failure items.

Environmental Characteristics. Stress is an environmental characteristic which is aroused by the presence of personal threatening conditions for the individual. General stress effects take on a distinct pattern relative to test anxiety. In conditions of high stress, LA subjects have been found to increase their level of functioning. Contrary to predictions that increased stress would actually serve to heighten anxiety symptoms, LA individuals exhibited greater effort and were more on task (Sarason, 1960). Consequently, performance improved and none of the associated dysfunctional thought processes commonly associated with more highly anxious individuals seem to appear (Deffenbacher, 1978). HA subjects respond quite the opposite to the general presence of stress. When tests are administered under conditions of high stress (Sarason, 1960), or normal stress, performance was poor. However, performance improved when level of stress was substantially reduced (Deffenbacher, 1978).

Audience presence has generally been found to decrease learning ability and deter performance for the population as a whole. However, recent research has shown that there is also a significant difference between high- and low-anxious individuals in regards to this variable (Ganzer, 1978). These studies have shown both positive (Geen, 1976) and negative correlations (Nicholl, 1984) of test anxiety with performance depending on particular conditions of audience presence.

"Evaluation emphasis" has received a great deal of attention in the literature as a particular type of stress. This topic refers to situations where examinees believe the test results will emphasize personal traits or abilities. Numerous studies have shown HA subjects perform in a comparable fashion to LA subjects when evaluation emphasis levels are low (Deffenbacher, 1978; Sattler, 1982; Wine, 1971). Conversely, highly evaluative circumstances lead to higher levels of performance for LA (Wine, 1971), and lower levels for HA (Sarason & Stoops, 1978).

Another type of environmental stress is caused by time pressure. Numerous studies reviewed by Matarazzo (1972) showed a decline in performance for highly anxious individuals on the same test depending on whether or not the test was administered in a timed vs. untimed administration. The LA group consistently exhibits no differences. Matarazzo also points out that studies which have shown the existence of a negative correlation between intelligence and level of test anxiety fail to take the aspect of timing into consideration, since only intelligence measures which are timed correlate in this fashion.

Test characteristics. Overall test difficulty is perhaps the most obvious test characteristic to differentially affect anxious individuals. Many of the first research attempts in this area used the Taylor Manifest Anxiety Scale. Highly anxious subjects were found to exhibit higher levels of performance than their less anxious counterparts on classical conditioning tasks. It was assumed level of anxiety was related to a state of "reactivity" or "excitability" that affected drive, which in turn was necessary for increased performance (Farber & Spence, 1953). When Sarason & Palola (1960) compared trait anxiety ratings with test anxiety ratings, the state anxiety ratings were found to be more accurate in predicting level of performance, especially when difficulty was taken into consideration.

There is overwhelming evidence for an interaction effect between test difficulty and level of anxiety (Child, 1954; Farber & Spence, 1953; Nicholls, 1984; Sarason, 1960). When HA students are given difficult problems, they are faced with a situation which they are unprepared to handle, and thus they are unable to complete the task at hand. But when the same individuals are faced with less difficult items, they can solve them with ease. The reverse is true for LA examinees. These persons are prepared to handle complex situations, and actually excel at them. However, should they be faced with test items of low difficulty, they will skim over them too quickly and consequently level of performance will decline. Weiner & Schneider (1971) explain this effect in terms of attribution theory. The low-anxious subject perceives failure as being due to a lack of effort. Thus initial failure experiences bring about greater effort. For high-anxious subjects, initial failure experiences are perceived to be caused by a lack of ability, with the result that all future performance deteriorates.

Test anxiety has also been found to influence person fit measures (Birnbaum, 1936; Schmitt & Crocker, 1984). Some subjects missed easier items, and correctly answered more difficult items. A good test should give equal results regardless of the order of the items on the test. However, Doris & Sarason (1955) found this not to be the case when comparing HA versus LA groups. They found patterns of success and failure differentially based upon level of anxiety on two forms of a test in which only the item orders differed. One possible explanation for this is found in the work of Galassi, Frierson, and Sharer (1981) in which they hypothesize that there are "critical moments" in a test which serve to increase test anxiety. Their research has found significant differences in the number of negative thoughts at different points in a test. They also tested for a specific anxiety by critical point interaction, but only found a tendency for significance ($p < .06$).

Initial failure experiences also interact with level of anxiety. It is interesting to note early studies of anxiety and performance used failure experiences to increase the level of anxiety. Later research continued to treat failure experiences as an independent variable. It was hypothesized that failure would elicit responses too strong for HA subjects to handle, and thus performance would decline. For LA individuals, some degree of failure should serve to increase motivational drive and subsequently improve overall performance (Child, 1954).

Perhaps test examiners would do well to take the advice of a researcher not even concerned with the effects of anxiety, but who clearly understood the possible negative effects of failure experiences. Hutt (1947) asserts that we do all individuals a disservice by having them encounter a series of failure experiences at the end of intelligence tests. Subsequent performance is likely to decrease and animosity develop towards the test administrator. HA subjects are likely to be overwhelmed and thus perform more poorly on subsequent sub-tests and low-anxious subjects are likely to decline even more rapidly than usual as they are experiencing a series of related failure tasks, the failure of each one increasing the probability of failure on the next.

Anxiety, Conventional Tests and Adaptive Tests

To date, the principle method utilized to demonstrate the efficacy of adaptive testing is to take an already administered conventional test, score it, compute item parameters, and then simulate an adaptive test administration using the responses already given on the conventional test. While this methodology has tended to demonstrate the accuracy of adaptive testing results as compared to conventional test results, little research has been conducted using actual adaptive tests. The following questions must be asked: Will the factors which currently affect conventional test results, such as level of test anxiety, show similar effects in adaptive situations? Will various methods of adaptive test administration differentially affect test outcome, whether or not test anxiety is controlled? And finally, will person fit measures be affected by anxiety and/or changes in item difficulty order?

Hypotheses

This study examined the varying effects of test anxiety on tests of different item orders. In this regard the following hypotheses were made: (1) the method of item administration can be manipulated so as to increase or decrease performance depending upon the individual's level of test anxiety, and (2) person fit statistics will deteriorate for conditions where there is an anxiety by difficulty order interaction.

Both hypotheses predict a difficulty order by anxiety interaction. When subjects are initially presented with items which are easy for them, low-anxious subjects are predicted to perform more poorly overall, as they do when presented with easy items in general. This change in effect is probably due to an insufficient level of effort caused by the appearance of lessened task importance, decreased stress, or a change in direction of attention. For high-anxious subjects in the same condition, the converse is thought to be the case. Easier items result in less stress, and fewer of the avoidance responses which ordinarily limit performance.

When subjects are initially presented with difficult items, the resultant effect on performance is predicted to be reversed for both anxiety groups. Low-anxious subjects will excel in the presence of items which challenge their ability level, resulting in on-task behaviors which increase their overall performance. However, when highly anxious subjects are continually administered items beyond their ability level, the level of stress will rise above their coping level, and dysfunctional thought processes will overwhelm the task relevant responding necessary for good test performance.

The first hypothesis can also be substantiated by comparing performance on a conventional test with performance on a test where item difficulty order has been manipulated. In this way a direct comparison can be made of the difference in performance which can be attributed to the effects of test anxiety and difficulty order versus the overall performance which is likely to be demonstrated in conventional test situations. The performance of low-anxious persons should be better on a test in which difficult items are initially presented as compared to high-anxious individuals whose performance should be worse in this situation.

Performance can also be evaluated on a within-test level. On a test where easy items are administered first, low-anxious subjects are likely to perform more poorly on the easy items than had they been presented later in the test, but performance should still improve once the test becomes more challenging. For high-anxious subjects in this situation, initial performance should be quite strong in the presence of easy items which are less stress inducing, but stress should increase, and performance deteriorate, once they begin to encounter the more difficult items. These effects should be further aggravated for individuals based upon their ability level. The performance of low-able, high-anxious subjects is more likely to deteriorate when they encounter the difficult section of the test, than are high-able subjects.

These within-test effects can also be conceived of in terms of the effect of time. The performance of high-anxious subjects should improve over time, as they get used to the test and their level of confidence begins to grow. But this effect will be moderated by item difficulty order and ability, which serve to impede or hasten the building of confidence. The performance of low-anxious subjects should deteriorate over time, as concentration for this group diminishes. But this effect will also be moderated by difficulty order and ability. Time will have less of an impact on low-anxious subjects encountering more difficult items later in the test, as interest will be increased by the new level of item difficulty. The low-anxious subjects encountering easy items at the end of the test will be affected by both the decreased interest associated with easier items as well as the negative influence of time. These effects will be further compounded by person ability. On a test where item difficulty declines over time, low-able subjects will encounter easier items later in the test than will more able subjects who may find the test items in the middle of the test to be easier for them.

When difficulty order is manipulated person fit statistics sometimes will indicate that a problem exists, particularly in the case of low-anxious subjects missing items which should be easy for them, and in the case of high-anxious subjects whose performance deteriorates following the administration of items which are too difficult for them. Fit is also likely to be affected by the influence of ability. Low-able subjects are likely to guess at more items as compared to high-able subjects, resulting in poorer fit statistics for all low-able subjects.

Method

Subjects

Subjects were drawn from the regular testing pool of the Johnson O'Connor Human Research Foundation (JOCRF). This non-profit testing organization annually administers aptitude tests to approximately ten thousand private clients in thirteen testing centers around the country. The foundation does not advertise its services, and thus referrals are generally made from previous clients, and often by other family members. For a fee of $450, six hours of group and individually administered aptitude tests are given. In addition, many of the clients complete a series of experimental tests. Following completion of the battery, subjects return for a 90-minute individual interpretive session with a test administrator.

The population seen at Johnson O'Connor ranges in age from fourteen to sixty. The 1233 subjects for this study were drawn from the regularly scheduled appointments in fifteen testing laboratories (two have since been closed).

Materials

Subjects were administered all of the regular aptitude tests during two half-day sessions which are usually given on the same day. The conventional vocabulary test was administered in two parts. The first part is a 20 item pre-test. The results of this pre-test are used to select the appropriate form of the conventional test. Neither administration was timed, nor was it supervised. It should be noted the group testing sessions are given in rooms with up to three other people. Timed tasks are given with the aid of an integrated audio-visual system which uses tapes and slides. All of the measures used in this study were administered untimed, usually at the end of one of the group sessions, after all of the timed tasks were completed.

The conventional test administered as part of the regular JOCRF battery is known as Worksample 690. According to the test's manual (Bowker, 1981), the test originated in the 1920's when the foundation first began to explore aptitudes. The 690 series is the seventh generation of vocabulary tests administered by the foundation. Worksample 695A, a 20-item placement test which uses items from the 690 series, is first administered and then scored. The 690 test consists of 225 items of which only 100-125 are administered depending upon the pre-test results. Worksample 690A uses items 1-125, 690B items 126-225, and 690C items 76-175. The test norms are computed based upon the scores of the entire national sample over the previous two years (N > 10,000) and have reliabilities in the range of .95-.97 (Bowker, 1981). The manual also cites numerous studies which examine the test's content, convergent, and criterion-related validities.

The experimental tests were constructed from 950 items found in the Foundation's item bank. Using Rasch modelling (Wright and Stone, 1979), many of the items had been placed on a common scale, providing a broad band item pool from which to draw. The item parameters for the 690 series as well as for many of the items in the item bank were computed using MSCALE (Wright, Rossner & Congdon, 1985). Only items which met rigorous inclusion criteria based upon infit (adherence to the overall model) and outfit (generally defined as a 'likelihood of guessing' measure) have been retained in the item bank. Contrary to the perspective of many proponents of the Rasch model, items with overly negative infit values have been retained for the purpose of this experiment, as the difficulty levels of these items is not in question. One hundred items were selected from the item bank for use in this study.

Three experimental tests were constructed. The first was designed to provide an alternate form for the 690 series. The goal for the Foundation was to construct a shorter test which encompassed the same difficulty range as the current test series. The items from the first experimental test were administered in increasing difficulty fashion, easy items first, most difficult items last. The other two tests differed from the first only in terms of difficulty order. In the second form, the more difficult items were administered initially, followed by the easier items. In the third form the items were distributed in random difficulty order. The same 100 items were used in all three cases.

Random assignment of test forms was accomplished by providing each testing laboratory with a single stack containing randomly ordered answer sheets for all three of the forms. Subjects were given the answer sheet which was on the top of the stack on the day they came in to take the test. A total of 650 answer sheets were distributed for the first form, and 325 for each of the other two forms. The discrepancy was based upon a requirement of the Foundation to have a large sample size in their test construction program. In the end, 1233 subjects completed the tests (619 the first form, 309 the second, and 305 the third).

Two personality tests were administered. Mandler and Sarason's Test Anxiety Scale (TAS) was administered in paper and pencil form. The test consists of a series of True-False items. In addition, a short series of test taking strategy items were administered.

All of the subjects also took a minimum of 18 aptitude tests (see Appendix 1 for descriptions). In general, these tests have high levels of reliability, and all have been correlated with scores on the 690 vocabulary series. Reliability information for all of the Johnson O'Connor standard battery tests can be found in Appendix 2.

A possible design problem existed in terms of lack of experimenter control. The use of an independent testing organization could present a series of unique problems, many of which would undoubtedly be realized only after data collection had commenced. These problems were limited due to the fact that the test administrators (TA's) are trained in an individually administered training course during a one month stay at the New York laboratory. When later changes are made to the battery, the TA's are informed through a constant news network of Test Information Bulletins and Technical Reports. In addition, the work of all TA's is reviewed annually by a computer analysis which compares mean test results obtained by each TA across all of their examinees, as compared to all of the TA's across the country. In this way a high degree of tester reliability is maintained.

Procedure

Three primary factors were included in the analyses: level of test ANXIETY, ABILITY and difficulty ORDER. The dependent variables included: the person MEASURE obtained on the experimental test, a measure of person fit -- INFIT, and the ABILITY RESIDUAL.

ANXIETY. The anxiety variable was computed as the raw score obtained on the Test Anxiety Scale. For the purpose of presenting tables, anxiety scores of all of those participating in the study were placed on a continuum with those scoring above .8 standard deviations from the sample mean labelled as HIGH anxious, those within the range ± .8 standard deviations of the sample mean as MEDIUM anxious, and those in the bottom group as LOW anxious. All of the actual analyses which utilized anxiety did so with anxiety entered as a continuous variable.

ABILITY. The ability measure for the conventional test is expressed in terms of a Vocabulary Scale Score (VSS). The VSS score was designed by the Jonnson O'Connor Research Foundation to act as a linear vocabulary scale which could be used to equate all vocabulary words and tests, regardless of ability. The Foundation has recently developed a formula which converts VSS scores to the log scale used by item response theory. Using this conversion formula, all ability measurements were able to be expressed on a common log scale in the computation of residual ability. Ability was expressed in VSS units when entered as a covariate in analysis of variance.

ORDER. Difficulty order was manipulated by randomly administering subjects one of three experimental tests: 1) the Easy-Hard form began with the easiest items and then proceeded to become more difficult, 2) the Hard-Easy form began with difficult items above the subject's ability level, or 3) the Random form where item difficulties fluctuated above and below the subject's ability level.

Item difficulties were established using the Rasch based measurement program known as MSCALE.

MEASURE. The person measure obtained on the experimental test was computed using MSCALE. The Rasch model centers items at zero, computes item difficulties, and then distributes persons according to their performance relative to the item difficulties. The measure obtained is also expressed in logits.

In order to analyze performance on the within-test level, ability measures were also computed for three different parts of each test: the first 30 items, the 40 middle items and the last 30 items. These three ability measures for each person were constructed by using the average item difficulties from all three test forms (effectively controlling for test length). A raw-score to logit conversion table was then generated for each group of items, within each test form, using a series of UCON iterations to compute the estimated measure given each possible raw score for that particular group of items (see Wright & Stone, 1976).

INFIT. Person-fit measures were also obtained using MSCALE. MSCALE calculates a *t* statistic to assess person fit called INFIT which is comparable between subjects. This statistic will remain small for those individuals whose response pattern is logical relative to their ability. However, the statistic will increase dramatically when numerous items are missed which are easier than the person's ability level, or when questions are answered correctly beyond the person's obtained ability score.

LOGIT CHANGE SCORE. The logit change score was computed as the difference between the score obtained on the conventional test and the score obtained on the experimental test where:

LOGIT CHANGE SCORE = Manipulated Performance - Conventional Performance

Poorer performance on the experimental test yields a negative logit change score, and better performance on the experimental test yields a positive logit change s
ore.

Results

The principle factors included in the analyses were all found to have some effect on performance, often interacting with the other variables in ways which had not been predicted. In general: a) test anxiety was found to be a factor in overall performance, logit change score, within-test performance, and person fit; b) item difficulty order was not a factor in person fit, but did play a role in overall performance, logit change score and within-test performance; and c) person ability was also a factor in within-test performance and person fit.

Predicting Overall Performance

Overall performance was analyzed using repeated measures analysis of variance (see Table 1). The dependent variable was the logit score obtained on the first 30 items, the middle 40 items, and the last 40 items on the experimental test. Difficulty order (Form) was entered as a fixed factor with three levels, and analyzed along with the continuous variable, test anxiety (TAS), and the covariate Ability (pre-test measure). As expected, there was a significant main effect for anxiety (see Figure 1) showing that performance decreases for th re anxious ($r = .27$). The main effect for difficulty order was also significant (see Figure 2) such that persons who took the items in the Easy-Hard order performed significantly better than did persons in the Hard-Easy or random conditions. While the interaction of Form by TAS was not significant using the repeated measures design, there are indications that a total score approach, rather than using repeated measures, would show otherwise (see Table 2).

As one would expect, the covariate of ability had a extremely strong relationship with the experimental test forms. However, there was also an indication that if one used a total score approcch there would be a significant item order by ability interaction (see Table 3) such that medium able persons perform better on the Easy-Hard form, while low and high able persons demonstrate no overall preference.

Within-Subjects Effects

The sub-test data were analyzed using a repeated measures design where the estimated ability at the beginning, middle and end of the test served as the dependent variables. Difficulty order, test anxiety and ability were all considered as factors. While most of the main effects and interactions of these factors were significant, the interesting information was found in the higher order interactions involving performance over time. In general, performance declined over time for the easy-hard form, improved for the hard-easy forms, and was unaffected on the random form. Test anxiety further moderated performance at different positions in the test, particularly for the Easy-Hard form. And finally, the greatest variance in performance taking into consideration was item position differences due to differences in ability level.

Within test performance varied for only two of the three forms when considering the item order by position interaction (see Figure 3). Performance decreased over time on the Easy-Hard form, and improved on the Hard-Easy form indicating a distinct preference for the easy items regardless of their position in the test. There were no significant differences for item position in the random order. And while the easy items were more likely to be answered correctly on the Easy-Hard form, performance on the difficult items was the same in both conditions.

Test anxiety was found to differentially affect within test performance depending on the item order. On the Easy-Hard form, low-anxious subjects peaked in performance in the middle of the test, as opposed to the high-anxious individuals who demonstrated their worse performance during the middle of the test (see Figure 4). Item position was not a factor in within-test performance for the medium anxious group. In addition, performance over time on the hard-easy and random forms were seemingly not affected by test anxiety, regardless of its intensity.

The majority of the within-test performance variance was due to the interaction of time with ability. Persons of low ability were found to perform worse during the middle of the test, as contrasted with persons of high ability whose performance peaks in the middle (see Figure 5). Performance of the medium-able group did not vary over time.

Item position was also a factor in a three-way interaction with difficulty order and ability (see Figure 6). The opposing performance which appeared for the high and low anxious groups on the Easy-Hard form was also apparent when looking at the Low and High ability groups with close to a .75 logit difference between performance in the middle versus the end of the form for the Low able group. Also, the High able group demonstrated a distinct preference for the Easy items regardless of item position.

Logit Change Score

The logit change score data was analyzed using multiple regression with the dependent variable Logit Change Score, defined as the experimental test measure minus the conventional test measure.

The significant main effect for Form showed that compared to performance on the conventional test, measured ability was better for people taking the Easy-Hard version of the experimental test, while it was worse for those taking the Hard-Easy form (see Figure 7, $\underline{F} = 6.8$, $\underline{p} < .001$). Test anxiety also played a role in predicting the Logit Change Score (see Figure 8, $\underline{F} = 15.7$, $\underline{p} < .001$). Low test anxious subjects received positive logit change scores. Performance was worse on their second test, as compared to high test anxious subjects whose performance was better the second time around. The interaction effect of Form by Test Anxiety was not significant.

Person Fit

Person fit was not affected by difficulty order, but was modified by the effects of test anxiety and person ability. Increases in test anxiety lead to a proportionately higher level of INFIT, $\underline{F} = 11.8$, $\underline{p} < .001$; $\underline{r} = .11$. However, since a higher INFIT actually indicates worse fit, the lower test anxious subjects are considered to be best fitting. The main effect for Form and the Form by TAS interaction were not significant.

When ability was added as a covariate the main effect of test anxiety continued to be significant, $\underline{F} = 14.7$, $\underline{p} < .001$. Ability, $\underline{F} = 11.8$, $\underline{p} < .001$, and the interaction of test anxiety with vocabulary ability (TAS X VSS), $\underline{F} = 16.4$, $\underline{p} < .001$, were also found to be significant. Correlational analysis confirmed the existence of a negative relationship, $\underline{r} = -.45$, between vocabulary ability and INFIT (see Figure 9). Low-able person are least likely to fit the item response theory model. This relationship seems to worsen even further for those of high test anxiety. However, there appears to be little or no difference in person fit for those of medium or high ability regardless of test anxiety.

Discussion

Item response theory assumes incorrectly that performance can be predicted based solely upon the fixed parameters of the items within a test. Contemporary proponents of the theory have neglected to consider that there is more to performance than just ability. Factors such as test anxiety and item order have been shown to affect performance. These variables cause test scores and within-test performance to increase for some, and decrease for others.

The predictions set forth in this study regarding person fit were confirmed, but only with regards to anxiety, and regardless of the item difficulty order. Further analysis also revealed that, contrary to the underlying assumptions of models of fit, INFIT is related to ability, fit worsens as ability declines. While current theories allow for this on short tests with poor equidiscrimination across the ability continuum, this should not be the case for longer tests.

As predicted, test anxiety affected overall performance; low test anxious subjects performed better overall than those with higher levels of anxiety. Anxiety also affects performance when repeated tests are administered. In comparison with the conventional test, scores for low-anxious subjects on the the experimental test were worse, while scores for high-anxious examinees were better. It appears that low test anxious individuals lose some of their competitive edge when facing their second test resulting in poorer performance. In contrast, high-anxious individuals use the initial testing session in order to build confidence and receive higher scores on the second test. This may also be an indication that the effects of a pretest diminish the differential effects of test anxiety.

The difficulty order was a significant factor for predicting total test score and in determining part of the difference in performance between the conventional and experimental test administrations. Subjects had better performance relative to the conventional test when given the Easy-Hard form, however their performance was worse on the Hard-Easy form, when the most difficult items were administered at the beginning of the test. This finding substantiates the predictions of Hutt that failure experiences decrease performance.

Within-test analyses revealed that while difficulty order is is a significant factor in overall performance, performance at different positions in the test is also affected by the form, level of anxiety, and ability. These effects seem to lessen greatly when items are administered in a random order. The Easy-Hard form is the most easily affected by individual differences. Low test anxious individuals improve in performance during the middle of the test, while the performance of high test anxious individuals declines. Within-test performance is further moderated by ability. Persons of high ability peak in their performance on the easy items, while low able persons actually peak on the most difficult items.

It is clear that items can no longer be viewed in a vacuum without regard for individual differences. The fundamental assumption of IRT that performance on any one item may be used to predict ability is wrong. Within-test performance, and performance on individual items, is subject to the effects of item order, test anxiety and ability, all interacting together.

The issue of changes in performance over time also has a major impact in determining test length. Adaptive testing is a commonly used method of shortening test length. Items are usually administered alternating between being above and below the examinee's ability level. This administration pattern is likely to produce performance similar to that found in the random difficulty order form. If this is the case, test anxiety and person ability is less likely to affect within-test performance. Test length appears to affect persons of differing abilities in unique ways. Depending upon the interaction of item order, anxiety and ability, some individuals seem to improve at later positions in the test, and others decline, at least up until a certain point when the trend reverses. In any such situation there must be a test length where performance stabilizes regardless of ability level. Unfortunately, it is the goal of most testing programs to decrease test length, irregardless of the needs of the examinee. It is clear that in order to maximize performance, test length could be individually tailored to the testing situation and the individual characteristics of the test taker.

This study demonstrated that item parameters cannot be fixed in advance without knowledge of personality factors and test characteristics. These results may be used by some to prove the non-utility of adaptive testing procedures. However, the findings could also be interpreted as demonstrating that particular individuals are better suited to unique test formats. Individual differences should be taken into consideration, and difficulty order and test length included as test parameters. The manipulation of these parameters within an adaptive testing situation will then maximize performance for all persons.

REFERENCES

Alpert, R. & Haber, R.N. Anxiety in academic achievement situations. Journal of Abnormal and Social Psychology, 1960, 61, 207-215.

Anastasi, A. Psychological Testing, New York: MacMillan, 1982.

Birnbaum, Menucha. Effect of Dissimulation Motivation and Anxiety on Response Pattern Appropriateness Measures. Applied Psychological Measurement, 10 (2), June 1986, pp. 167-174.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F. & Novick, R., Statistical Theories of Mental Test Scores. Massachusetts: Addison-Wesley, 1968.

Bowker, R. English Vocabulary Manual. Chicago: Johnson O'Connor Research Foundation, 1981.

Child, I.L. Personality. Annual Review of Psychology, 1954, 5, 149-171.

Daniel, M. Reliabilities and Standard Errors of Laboratory Worksamples. Test Information Bulletin 1980-7. Chicago:Johnson O'Connor Research Foundation, 1980.

Daniel, M. Technical Report 1983-6. Chicago:Johnson O'Connor Research Foundation, 1983.

Deffenbacher, J.L. Relationship of worry and emotionality to performance on the Miller Analogies Test. Journal of Educational Psychology, 1977, 69, 191-195.

Deffenbacher, J.L. Worry, emotionality, and task-generated interference in test anxiety: an empirical test of attentional theory. Journal of Educational Psychology, 1978, 70, 248-254.

Doctor, R.M. & Altman, F. Worry and emotionality as components of test anxiety: replication and further data. Psychological Reports, 1969, 24, 563-568.

Doris, J. & Sarason, S. Test anxiety and blame assignment in a failure situation. Journal of Abnormal and Social Psychology, 1955, 50, 335-338.

Eysenck, Michael (1979). Anxiety, Learning, and Memory: A reconceptualization. Journal of Research in Personality, 13, 363-385.

Galassi, J.P., Frierson, H.T. & Sharer, R. Behavior of high, moderate and low test anxious students during an actual test situation. Journal of Consulting and Clinical Psychology, 1981, 51-62.

Ganzer, V.J. Effects of audience presence and test anxiety on learning and retention in a serial learning situation. Journal of Personality and Social Psychology, 1968, 8, 194-199.

Geen, R.G. Test anxiety, observation and range of cue utilization. British Journal of Social and Clinical Psychology, 1976, 15, 253-259.

Hambleton, Ronald K. & Van der Linden, Wm J. Advances in item response theory and applications: An introduction. Applied Psychological Measurement, Vol. 6, No. 4, Fall 1982, 373-378.

Hollandsworth, J.G., Glazeski, R.C., Kirkland, K., Jones, G.E., and VanNorman, L.R. An analyses of the nature and effects of test anxiety: cognitive, behavioral, and physiological components. Cognitive Therapy and Research, 1979, 3, 165-180.

Hutt, M.L. A Clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

Kingston, N.M. & Dorans, N.J. Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, Vol. 8, No. 2, Spring 1984, 147-154.

Kirkland, M. Effects of Tests on students and schools. Review of Educational Research, 1971, 41, 318-319.

Leon, M.R. & Revelle, W. The effects of anxiety on analogical reasoning: a test of three models. Journal of Personality and Social Psychology, 49, 1302-1315.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. New Jersey: Lawrence Erlbaum Associates, 1980.

Lord, F.M. Practical considerations of item characteristic curve theory. Journal of Educational Measurement, Vol. 14, No. 2, Summer 1977, 117-138.

Mandler, G. & Sarason, S.B. A study of anxiety and learning. Journal of Abnormal and Social Psychology, 1952, 47, 166-173.

Mann, J. Vicarious desensitization of test anxiety through observation of videotaped treatment. Journal of Counseling Psychology, 1972, 9, 1-7.

Matarazzo, J.D. Wechsler's Measurement and Appraisal of Adult Intelligence, New York: Oxford University Press, 1972.

McBride, J.R. Bandwidth, fidelity, and adaptive tests. InCATC-2 1975. Atlanata, GA: Atlanta Public Schools, 1976.

Meichenbaum, D. Cognitive modification of test anxious college students. Journal of Consulting and Clinical Psychology, 1972, 39, 370-380.

Meunier, C. & Rule, B.G. Anxiety, Confidence and Conformity. Journal of Personality, 1969, 35, 498-504.

Mitchell, K.R. & Ingham, R.J. The effects of general anxiety on group desensitization of test anxiety. Behavior Research and Therapy, 1970, 8, 69-78.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. Relationship between corresponding armed services vocational aptitude battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, Vol. 8, No. 2, Spring 1984, 155-164.

Morris, L.W. & Liebert, R.M. Effects of anxiety on timed and untimed intelligence tests. Journal of Consulting and Clinical Psychology, 1969, 33, 240-244.

Nicholls, J.G. Achievement motivation: conceptions of ability, subjective experience, task choice, and performance. Psychological Review, 1984, 91, 328.

Paul. G. & Eriksen, C.W. Effects of anxiety on "real life" examinations. Journal of Personality, 1964, 32, 480-494.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Ruebush, B.K. Interfering and facilitating effects of test anxiety. Journal of Abnormal and Social Psychology, 1960, 33(2), 205-212.

Sarason, I.G. The effects of anxiety, reassurance, and meaningfulness of material to be learned in verbal learning. Journal of Experimental Psychology, 1958, 56, 472-477.

Sarason, I.G. Empirical findings and theoretical problems in the use of anxiety scales. Psychological Bulletin, 1960, 57, 403-415.

Sarason, I.G. & Palola, E.G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. Journal of Experimental Psychology, 1960, 59, 185-191.

Sarason, S.B. & Stoops, R. Test anxiety and the passage of time. Journal of Consulting and Clinical Psychology, 1978, 46, 102-109.

Sattler, J.M. Assessment of Children's Intelligence and Special Abilities, Boston: Allyn and Bacon, Inc., 1982.

Schmitt, A.P. & Crocker, L. (1984). The relationship between test anxiety and person fit measures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Spielberger, C.D. & Katzenmeyer, W.G. Manifest anxiety, intelligence, and college grades. Journal of Consulting Psychology, 1959, 23, 278.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. Applied Psychological Measurement. Vol 7, No. 2, Spring 1983, 211-226.

Trabin, T & Weiss, D. The person response curve: fit of individuals to item response theory models. New Horizons in Testing. New York: Academic Press, Inc., 1983.

Trapp, E. & Kausler, D. Test anxiety and goal setting behavior.Journal of Consulting Psychology, 1958, 22, 31-34.

Urry, V.W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, Vol. 14, No. 2, Summer 1977, 181-196.

Weiner, B. & Schneider, K. Drive versus cognitive theory: a reply to Boor and Harmon. Journal of Personality and Social Psychology, 18, 258-262.

Wine, J. Test anxiety and direction of attention. Psychological Bulletin, 1971, 76, 92-104.

Wood, R.L. Wingersky, M.S., & Lord, F.M.(1976). LOGIST: A Computer Program for estimating examinee ability and item parameters (RM-76-6). Princeton NJ: Educational Testing Services.

Wright, B.D. & Mead, R.J. BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1976.

Wright, B.D., Rossner, M. & Congdon, R. MSCALE, Statistical Laboratory, Department of Education, University of Chicago, 1985.

Wright, B.D. & Stone, M.H. Best Test Design. Chicago: MESA Press, 1976.

Yerkes, R.M. & Dodson, J.D. The relation of strength of stimuli torapidity of habit-formation. Journal of Comparative Neurology and Psychology, 18,459-482.

Table 1
Analysis of Variance

Source	SS	df	MS	F
Covariate				
Ability (Pre-test)	393.30	1	393.30	774.94***
Between Subjects				
(Item) Order	3.13	2	1.57	3.08*
Anxiety	2.97	1	2.97	5.86*
Order X Ability	2.61	2	1.31	2.57
Order X TAS	2.72	2	1.36	2.68
Ability X TAS	2.84	1	2.84	5.60*
Order X Able X TAS	2.65	2	1.33	2.61
Error	597.36	1177	0.51	
Within Subjects				
Position,lin	0.20	1	0.23	0.55
Position,quad	2.62	1	2.62	10.96***
Ability X Pos,lin	0.21	1	0.21	0.56
Ability X Pos,quad	2.48	1	2.48	10.36***
Order X Pos,lin	5.79	2	2.90	7.84***
Order X Pos,quad	0.53	2	0.27	1.11
TAS X Pos,lin	0.00	1	0.00	0.01
TAS X Pos,quad	0.44	1	0.44	1.83
Order X Able X Pos,lin	6.71	2	3.35	9.07***
Order X Able X Pos,quad	0.22	2	0.11	0.47
Order X TAS X Pos,lin	0.61	2	0.30	0.82
Order X TAS X Pos,quad	1.59	2	0.79	3.32*
Ability X TAS X Pos,lin	0.00	1	0.00	0.01
Ability X TAS X Pos,quad	0.66	1	0.66	2.76
Order X TAS X Able X Pos,lin	0.59	2	0.30	0.80
Order X TAS X Able X Pos,qu	1.02	2	0.51	2.12
Error,lin	435.23	1177	0.37	
Error,quad	281.81	1177	0.24	

Pos=Position
Lin=Linear effect for Position
Quad=Quadratic effect for position

* $p < .05$
** $p < .01$
*** $p < .001$

Table 2

Performance as a Function of Form and Anxiety

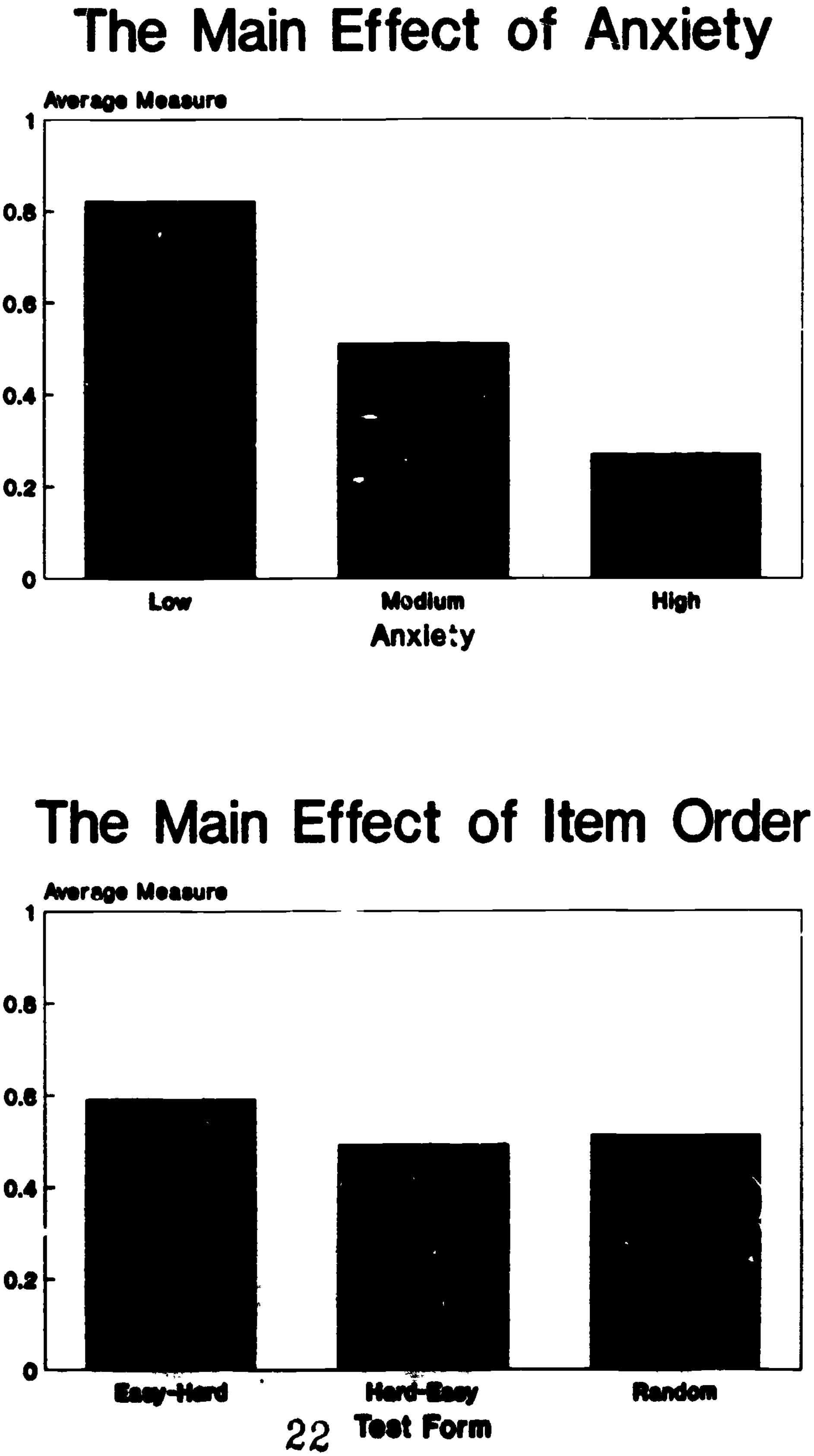
Form	Low Anxious	Medium Anxious	High Anxious
Easy-Hard	.88	.57	.27
Hard-Easy	.71	.43	.31
Random	.82	.44	.22

Table 3

Performance as a Function of Form & Pre-Test Ability

Form	Low Able	Medium Able	High Able
Easy-Hard	-.50	.54	1.64
Hard-Easy	-.53	.39	1.51
Random	-.54	.48	1.60

Figures 1 & 2

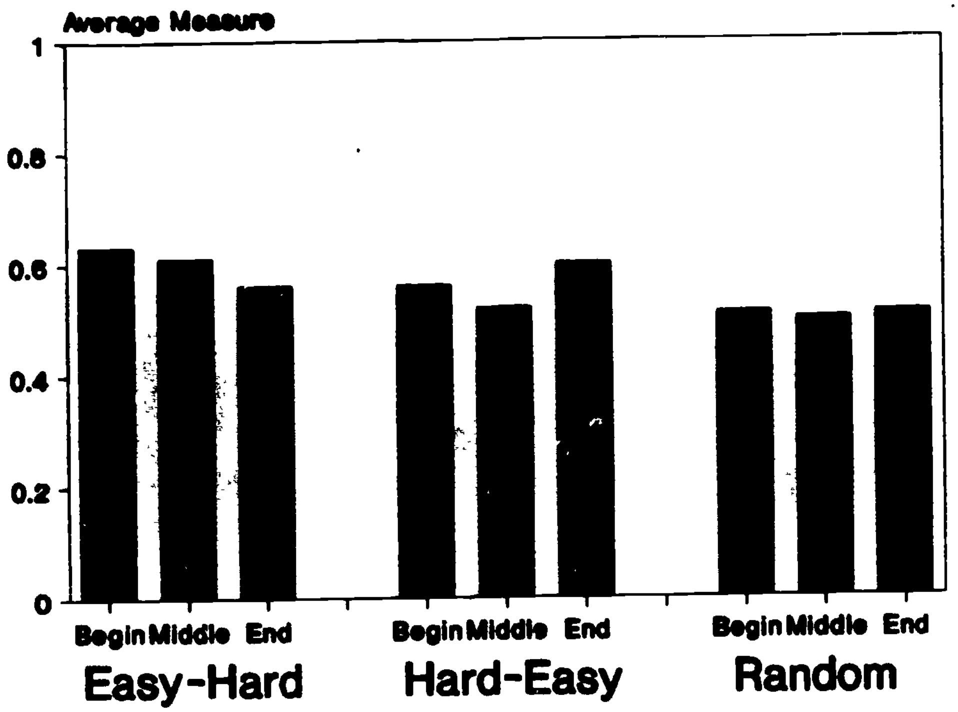


The Main Effect of Anxiety

The Main Effect of Item Order

Figures 3 & 5

Item Order X Position



Ability X Position

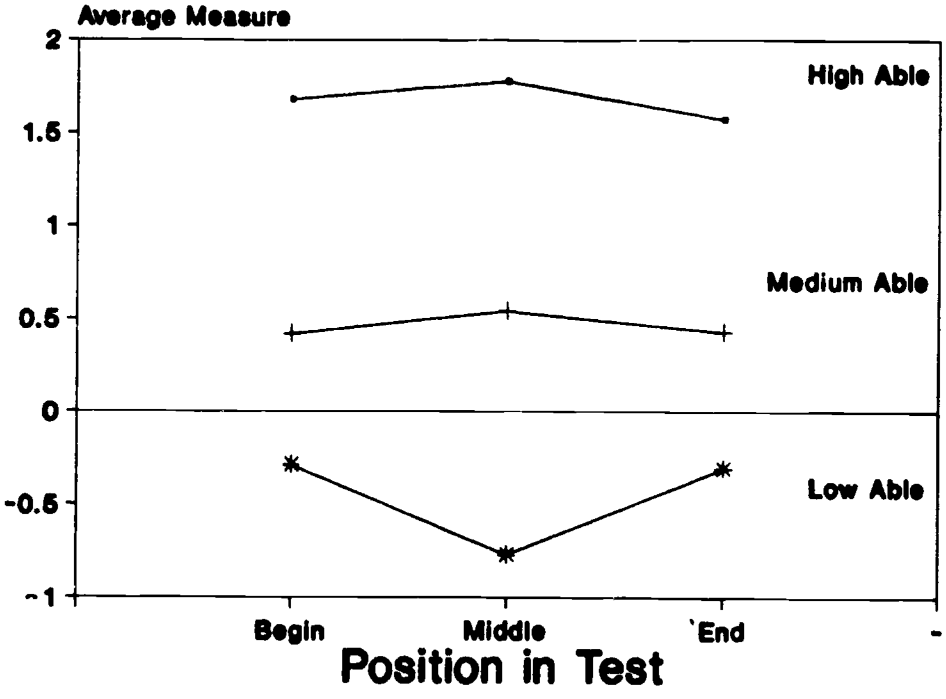
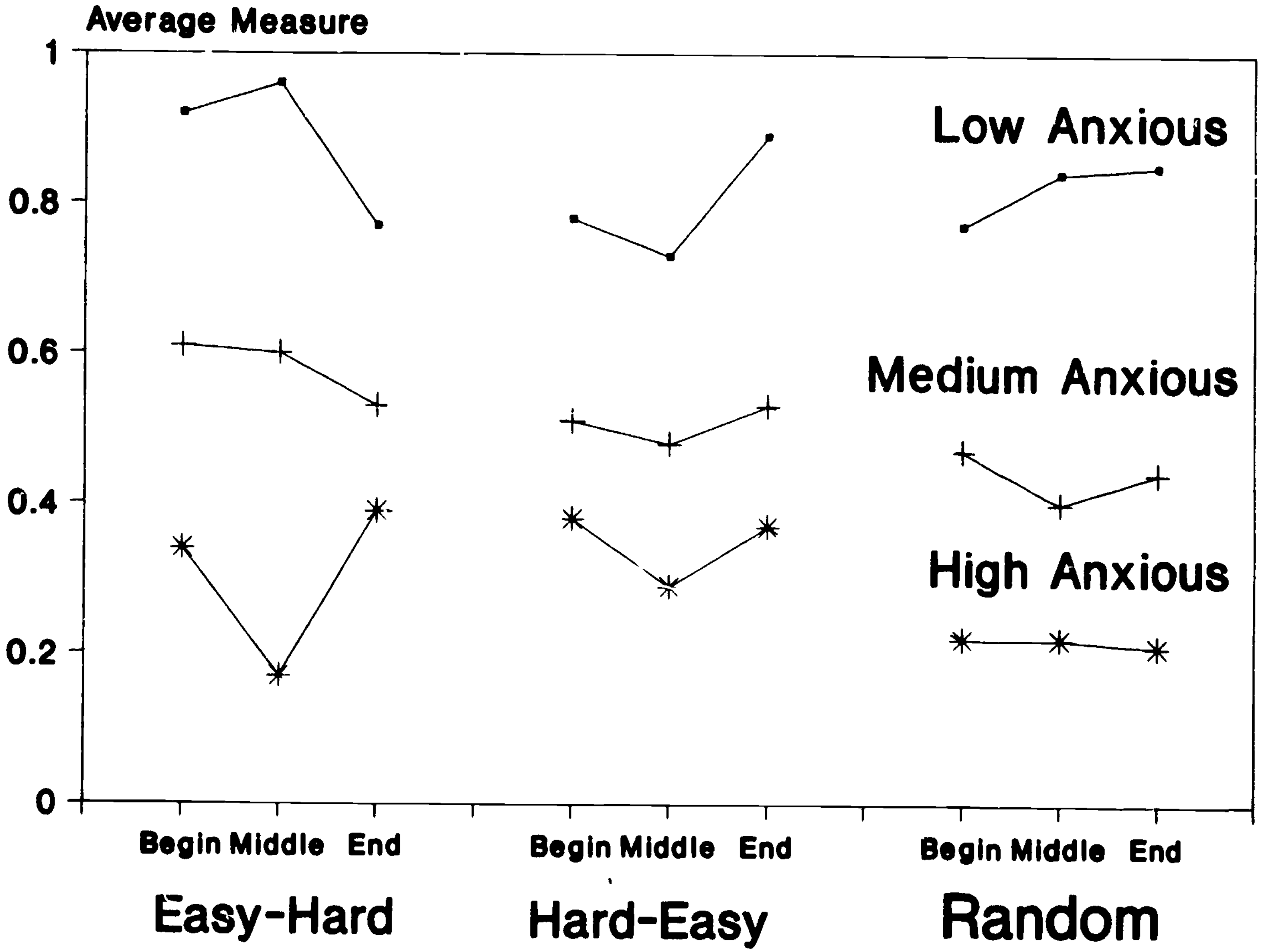


Figure 4

Item Order X Anxiety



Item Order X Pre-Tested Ability

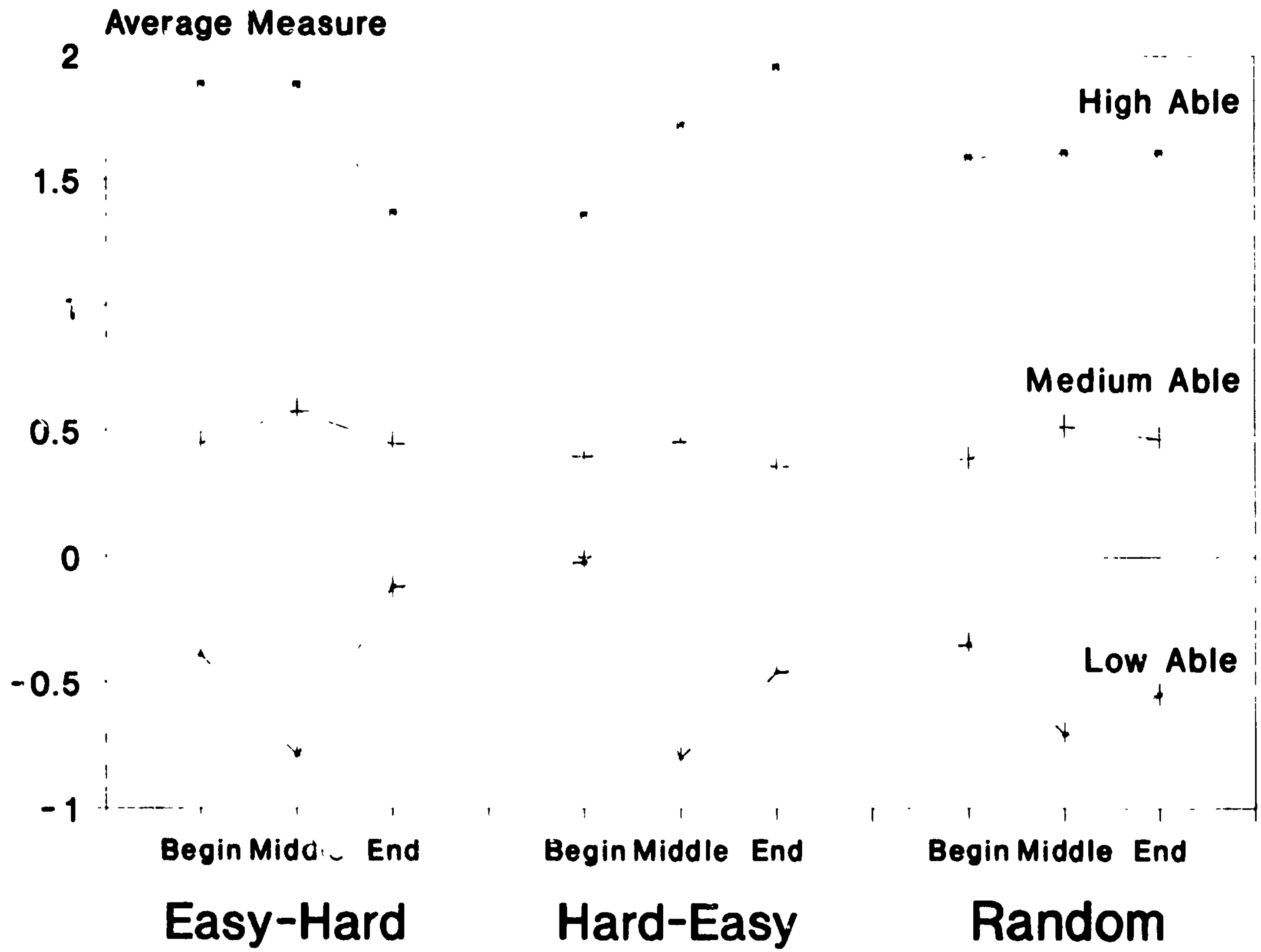
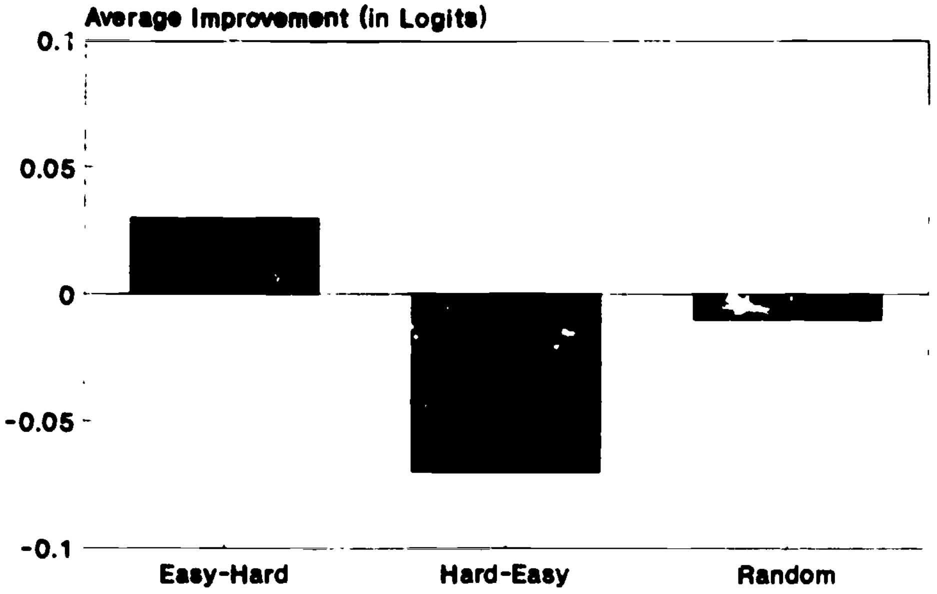


Figure 6

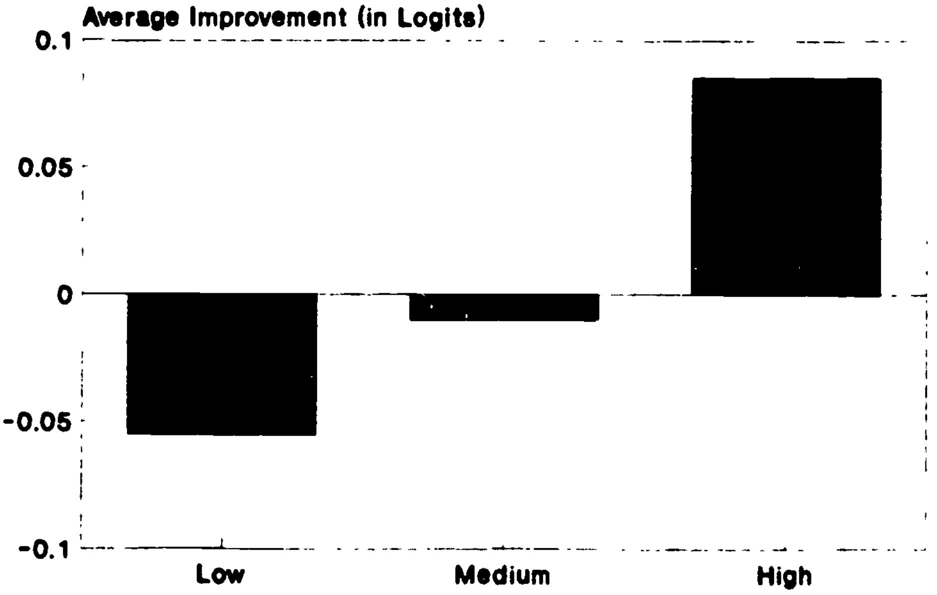
Figures 7 & 8

Comparing Differences Between
Pre-Test and Experimental Test



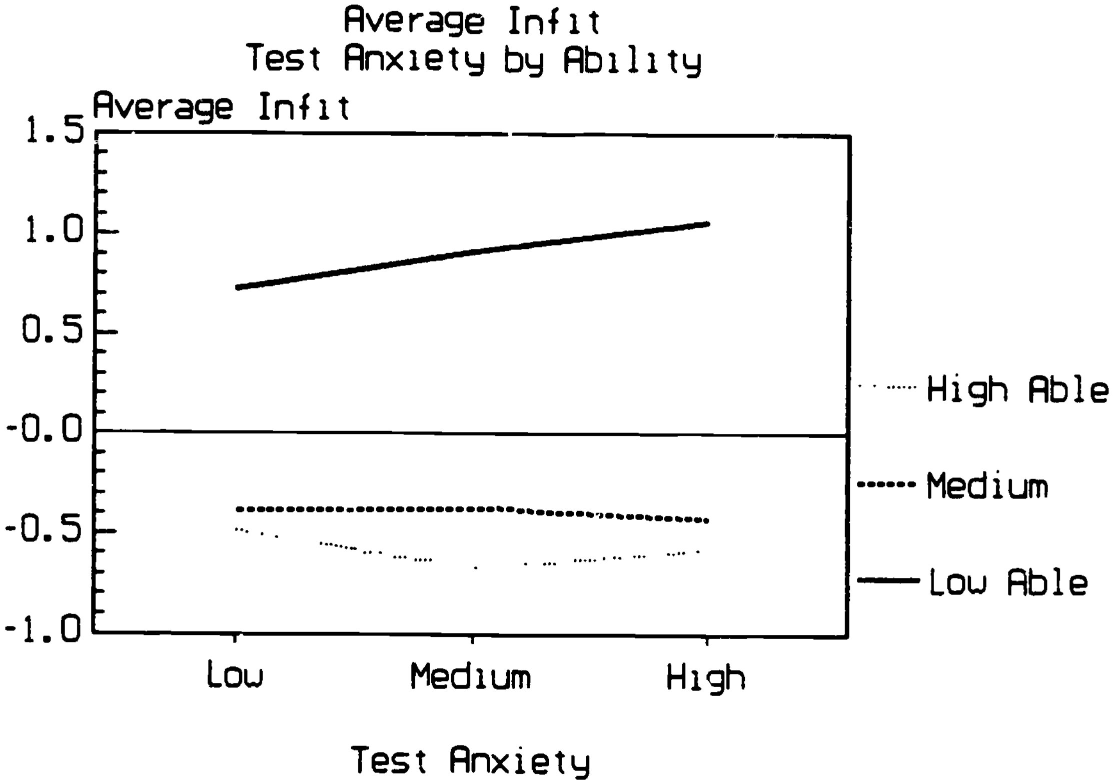
By Form

Comparing Differences Between
Pre-Test and Experimental Test



By Anxiety

Figure 9


Average Infit
Test Anxiety by Ability
Average Infit
1.5
1.0
0.5
-0.0
-0.5
-1.0
Low
Medium
High
Test Anxiety
High Able
Medium
Low Able