1991

Effects of test administrations on general, test, and computer anxiety, and efficacy measures

Susan Kiskis

EFFECTS OF TEST ADMINISTRATIONS ON GENERAL, TEST, AND COMPUTER ANXIETY, AND EFFICACY MEASURES

A Thesis

Presented to the

Faculty of

California State

University, San Bernardino

---

by

Susan Kiskis

March 1991

Approved:

[redacted]

Richard Trafton, Ph.D. Chair — 3/14/91 Date

[redacted]

Matt Riggs, Ph.D.

[redacted]

Diane Halpern, Ph.D.

EFFECTS OF TEST ADMINISTRATIONS ON GENERAL, TEST, AND COMPUTER ANXIETY, AND EFFICACY MEASURES

A Thesis

Presented to the

Faculty of

California State

University, San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Psychology

---

by

Susan Kiskis

March 1991

ABSTRACT

Using Item Response Theory, a computer adaptive test of clerical abilities was developed and compared to conventional computer and paper-and-pencil test administrations. Tests were administered to 150 applicants at a personnel agency. The study was designed to demonstrate the equivalency of computer adaptive testing to other test conditions using general and specific measures of anxiety and efficacy. It was concluded that computer adaptive testing was not only a psychometrically sound means of assessment, but also posed no threat to anxiety in evaluative situations. Conclusions regarding the impact of these findings are discussed.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Personnel selection processes typically consist of standardized paper-and-pencil tests. This method has proven to be time-consuming and costly. With the advent of computer testing technology and development of modern test theory, computer-administered and scored instruments have begun to be prevalent in personnel selection and screening. Applicants are increasingly required to use computers, both in the workplace and during the screening process. Computer testing technology has reduced test administration costs, relieved test administrators of the burdensome and time-consuming process of hand-grading and scoring, and allowed for multiple, concurrent test administrations.

Employing computerized tests presents an opportunity for fundamental changes in psychometric evaluation of tests. Even though computers offer advantages, controversial issues related to the effects on the examinee remain important for research.

Bracker and Pearson (1986) predicted that more than thirty-five million workers will be affected by growing technological changes. They will be faced with a the need to master complex computer hardware and software. A legitimate question arises regarding aptitude and ability.

It is important to understand how these might affect the performance of examinees under computer-administered conditions, especially if computer anxiety were present. Additionally, applied research in this area is lacking.

Most test and computer anxiety research focuses on educational settings. Research in applied settings such as in personnel selection and screening is needed. Continued research to answer questions regarding computer testing and computer anxiety must be conducted to assist American business as it invests in this new technology.

Aside from conventional computer-administered tests, modern test theory has begun to provide a firm foundation for development of personnel testing strategies and methods. Research by Weiss (1983) and Lord (1980) has led to tailored and branched testing, as well as multistage tests that create a hierarchy of subtests for each examinee. These "pyramidal" tests allow each examinee to begin with the same item, and then branch out to more difficult or easy items, depending on the response.

Limitations found with pyramidal tests guided the development of "stradaptive" tests (Vale & Weiss, 1978; Weiss, 1983). This method allows the examinee to proceed through several subtests or strata. As correct or incorrect responses are made, the test branches to more difficult or easier strata. This resulted in a more

individualized testing procedure.

"Adaptive" or tailored tests use statistical calculations to match examinee performance on each item. Relatively easier items are administered to a low ability examinees; more difficult items are administered to high ability examinees. In this way, the test items discriminate more accurately among low or high examinees. Lord (1950) was instrumental in the eventual development of the adaptive testing methods.

These new testing methods are readily applied in the workplace due to innovative computer technologies. Test construction has been simplified with computer software packages that open the door to adaptive testing.

Few studies currently exist that establish and encourage the use of computer-adaptive testing in an organizational setting. No research was found that looked at the combined effects of general anxiety, test anxiety, and computer anxiety when test procedures included computer-adaptive methods.

Personnel screening/selection methods do not take examinee artifacts such as general, test, or computer anxiety into account. Yet Wine (1971) found that selection batteries administered under highly evaluative conditions increased anxiety. Ample evidence for the deleterious effects of anxiety in evaluative situations has been documented by many researchers (Sarason, 1960,

1973; Wine, 1971; Rosen, Sears, and Weil, 1987a). Ignoring this important part of the applicant testing procedure may have profound, long-term effects on the predicted productivity of workers.

GENERAL AND TEST ANXIETY

Conceptually, general anxiety has been found to be a coping mechanism that pervades our lives, one that contributes to our survival as the fail-safe mechanism for adaptation. In non-adaptive modes, it promotes incompetence and misery (Sieber, 1978). Anxiety responses may vary, depending on past experiences, individual coping responses, problem context, and the level of anxiety evoked (Sarason, 1978).

Non-adaptive modes of anxiety include emotional responses such as panic, worry, anger, shame, or the desire to escape. Anxiety, as a general system of conditions, depends on the nature of the associated demands, feedback, or prior learning. Spielberger (1966) introduced the state-trait model of anxiety, which included these conditions. The transitory "State" anxiety occurred when perceived stimuli (real or imagined) resulted in emotional and behavioral responses. Trait anxiety was found to be a more stable personality characteristic.

Spielberger (1966) defined State anxiety as a reaction of heightened arousal, vigilance, enthusiasm, worry and fear, confusion, anger, lowered self-esteem, and other negative conditions. Prior experience moderates the

level of the State anxiety reaction. Cognitive reappraisal acts to guide State anxiety responses. Maladaptive coping strategies led to high anxiety as a disruptive emotional reaction, along with less constructive ways to deal with the immediate problem.

Early development of more specific measures of anxiety began with the Test Anxiety Scale (TAS; Sarason, 1952). It measured need for achievement, hostility, and lack of protection measures. The study compared predictive utility of Spielberger's General Anxiety scale to the TAS and found a more consistent correlation with performance with the TAS. Sarason concluded that the TAS would be useful as an index of proneness to performance disruptions. He emphasized the need to consider test anxiety when interpreting intellectual performance.

Sarason (1961) found that a more specific anxiety scale resulted in better prediction. The closer the scale content in relation to the assessed performance, the better the prediction.

Sarason's (1960) earlier model of test anxiety presented both cognitive and somatic components. Results of this study found that high test anxious individuals interpreted tests as serious threats to well-being with emotional reactivity. High test anxious subjects had increased sensitivity to cues which suggested the

imminence of an evaluative situation.

Evaluational stressors, or achievement-orienting instructions, were found to increase test anxiety (Sarason, 1978). Using the TAS, subjects were told that the task to be completed was a measure of intelligence. Performance was deleteriously affected by these instructions.

Wrightsman (1962) confirmed the "emotional reactivity" component of test anxiety. In his research, college students were told that results of intelligence tests were either important or unimportant. Performance decrements increased with emphasis on the importance of the test results. He concluded that a stressful situation interfered with successful performance of high anxious subjects. These results also underscored the strong influence of the testing instructions.

Wine (1971) reviewed the literature and found emphasis on the relationship between cognitive rumination (worry) and test performance. Worry and emotionality were viewed as major components of test anxiety in a study by Liebert and Morris (1967). Using their own worry-emotionality scale, they concluded that the cognitive, self-preoccupied component of test anxiety clearly interfered with performance.

Test anxiety has been strongly associated with a class of stimuli related to past evaluative or testing

experiences (Wine, 1971). These stimuli included test stimuli, interpretation of test stimuli, State anxiety reactions, cognitive reappraisal, coping, avoidance, and defensiveness (Spielberger, 1966).

Sarason (1957, 1963) found fear of failure in evaluative situations was based on negative interpretations of past experiences as opposed to fear of failing to carry out operations required at that specific time. Sarason found evidence that it acts as a non-intellectual influence.

In 1978, Spielberger used the TAS to develop the Test Anxiety Inventory (TAI). Factor analysis of the TAS and the TAI resulted in high factor loadings for worry and emotionality. He used these factors to define test anxiety. Spielberger concluded that the TAI could be used as an assessment device to determine the "situation-specific personality trait" of test anxiety.

Sieber (1978) devised questionnaires and surveys to assess the validity of test anxiety measures by Spielberger, Gorsuch, and Lushene (1970) and Sarason (1973). He found that the self-report scales proved to be the most valid predictors of test anxiety in both studies.

Additional studies on test anxiety point to the task-irrelevant responses in the testing situations as the key to performance interference. Two studies (Mandler and

Sarason, 1952; Wine, 1971) presented interference models of test anxiety. Wine's model focused on "evaluational apprehension," as a form of prior negative experience. He found that high test anxious subjects reacted with cognitive concern and subsequent decreased performance. The cognitive concern was viewed as an interfering anxiety that led to negative self-reflection. It was also found to compete with an individual's ability to perform.

Culler and Holahan (1980) conducted a study of the Wine (1971) and the Mandler and Sarason (1952) models. They found significant decrements in performance for high test anxious individuals. This effect was found to be mediated by increased study time, with high test anxious subjects having poorer study habits.

Anxious self-preoccupation and self-focusing remained the primary factors of current test anxiety theory. Researchers continued to move away from the autonomic response model which suggested anxiety was a result of maladaptive levels of autonomic arousal (Holroyd, Westbrook, Woolf, and Badhorn, 1978). In their study, both high and low anxious subjects experienced virtually identical changes in autonomic responses, such as heart rate and skin conductance, as a result of an evaluative condition. This study provided evidence for the cognitive interference models of test anxiety. Deficits in information processing did not result from maladaptive

levels of autonomic arousal. Conceptualization of test anxiety as a cognitive phenomena was supported in subsequent research. Additionally, the interference models of test anxiety related more strongly to distractions in attention than to autonomic responding.

COMPUTER ANXIETY

The growing trend to computerize test administrations has led to increased research to assess effects of test anxiety and computer anxiety. Similar performance decrements were found with both computer anxiety and test anxiety. Individuals found to be high test anxious have also been found to be high computer anxious (Rubin, 1981). Not only does the threat of evaluation interfere with performance in testing environments, but subjects experienced computer anxiety in a computer-administered testing setting.

Weinberg (1982) estimated that nearly 33% of the people sampled in a study of 500 college students and corporate managers could be categorized as "cyberphobic" or having high computer anxiety. Paul (1982) also found 30% of the business community as having experienced some form of anxiety about computers in a survey he conducted. It may be expected that these workers would bring a certain level of anxiety with them into the work environment, thereby interfering with their performance under testing conditions.

Development of scales to measure effects of computer anxiety began with strong emphasis on attitudes toward computers and included measures of task-specific efficacy.

Hill, Smith, and Mann (1987) studied computer anxiety using a measure of self-efficacy, perceived complexity of innovations and cognitive laziness to assess attitudes associated with the use of computers. They labeled computer anxiety as "technophobia." Subjects who had a high sense of efficacy regarding use of computers were found to be more likely to use them. Their study underscored the importance of efficacy beliefs in the decision to adopt an innovation.

Morrow, Prell, and McElroy (1986) outlined correlates of computer anxiety in their study of college students. They found that behavioral correlates, such as prior experience with computers and computer knowledge, accounted for 36% of the variance in computer anxiety. Self-reported behaviors explained more of the variance in computer anxiety than personality or attitudinal correlates. They concluded that computer anxiety may be a function of prior experience and viewed it as a modifiable condition.

Developing another label of computer anxiety, Brod (1982) proposed that the condition of "technostress" resulted from an inability to adapt to the introduction and operation of new technology. Correlates affecting this included age, prior computer experience, and perceived control over new tasks. In an applied setting,

Brod found that control of potentially negative statements about oneself, viewed in this context as self-efficacy, greatly improved the coping mechanisms of cyberphobics.

O'Neil and Richardson (1978) presented a review of studies in which modifications of instructional procedures in computer learning environments were attempted to reduce computer anxiety. However, as Hedl and O'Neil (1977) also found control over instructional variables and immediate feedback in the computer environment led to increased levels of anxiety. Instructions prior to the test led to significant differences between groups when an evaluative threat was introduced. Similar conclusions were found by Wrightsman (1961).

Rosen, Sears, and Weil (1987b) generated a 3-scale measure for computer anxiety comprised of attitudinal, anxiety, and cognitive elements. Elements of resistance to computers, fear or anxiety towards computers, and hostile or aggressive thoughts about computers were researched in a 5-part study using university students. The first two studies included use of the STAI.

Rosen's scales consisted of a computer anxiety rating scale (CARS; Rosen, 1988), attitudes toward computers scale (ATCS; Rosen, 1987), and a computer experience demographics questionnaire (CEDQ; Rosen, 1987). A negative relation was found between the anxiety and attitude scales ($r = -.29$, $p < .01$). Subjects with prior

computer experience and positive attitudes toward computers were the least anxious. "General computer anxiety--operating the machines themselves" was found to account for 40.3% of the variance in computer anxiety. Women were found to have a less positive attitude overall towards computers, with the greatest fear that improper use of the computer would result in damage. Of the total sample, 14% were high anxious. Graduate students served as subjects and had a mean age of 37 years.

In a second study, Rosen, et al. (1987a) included the STAI. State anxiety was negatively correlated with computer attitudes ($r = -.17$, $p < .01$). The sample for this study consisted of 66% non-whites, such as blacks and Hispanics.

BENEFITS OF COMPUTERS

Most research has been conducted on the effects of conventional, computer-administered testing. The argument in the forefront states that computerized psychological testing is depersonalizing. In clinical research, studies typically have cast the client as an object of automated manipulations by a computer.

Burke and Normand (1987) reviewed the literature and found overwhelming evidence that clients reacted favorably to computerized testing. Investigating accuracy of information about sensitive areas, such as use of alcohol, drugs, and tobacco, Skinner and Allen (1983) had subjects answer questionnaires in face-to-face, self-report, and computer conditions. The computer condition was rated as most relaxing and interesting. The study was conducted on a clinical patient population and was therefore difficult to generalize to other settings. However, their research called into question the theory that computers were depersonalizing. It also underscored the need for further research and a redefinition of the interactions and effects of computers on human subjects.

Hulin, Drasgow, and Parsons (1983) underscored the increased measurement accuracy available with computerized testing, along with reduced boredom, fatigue and testing

time. Test administration becomes more flexible since it could be conducted at any time with scheduling of test administrators becoming unnecessary. Urry (1977) found substantial cost reductions with computerized tests. Finally, there is an improved scoring efficiency and accuracy.

These studies provide evidence for the changing view toward computers, at least on a theoretical level. Research has begun to reveal the user-friendly aspects of computers; users find them appealing in situations where sensitive information is to be divulged. Aside from the increased accuracy in testing and scoring, reduced test-taking time, reduced costs involved with testing, there remains a need to continue research to show clear evidence for the human factor in computer testing.

SELF-EFFICACY

Using the framework of self-efficacy theory (Bandura, 1977), efficacy perceptions are defined as the belief of one's ability to accomplish a particular task. Self-efficacy has been found to influence the choice to engage in a task, along with the effort and persistence necessary to perform it. Bandura established that perceived self-efficacy for particular task increased if prior experience provided positive information about related competencies for that task.

Bandura (1977) described persons high in self-efficacy as those able and personally effective. High test anxious persons were found to have low self-efficacy due to preoccupation with fear of failure and self-blame. Perceived positive self-efficacy functions as a cognitive setting for success; self-blame on the other hand, may be viewed as the converse of efficacy.

Consistent with Bandura's claim that self-efficacy directly affects levels of task performance, Wood and Locke (1987) found a significant positive relationship between efficacy and academic performance in their study of college students ($r = .27$, $p < .01$).

When Miura (1987) used perceived competency measures as an index of computer anxiety, women were found to be

more often high test anxious. His study focused on self-efficacy judgment as a belief that one could successfully execute a particular course of behavior. Using college students, Miura assessed self-efficacy beliefs as related to ownership, use, programming, and decision to take future coursework in computers. He found a relation between self-efficacy and current and past enrollment in school computer programming courses (r = .29, .47, and .31 respectively, $p < .001$). Factor analysis of the results generated the highest factor loadings for prior experience.

Self-preoccupation with perceptions of being "unable" lead to attentional misdirection. High anxious individuals are unable to focus on the task at hand. Sarason (1978) found this negative preoccupation to be at the core of test anxiety.

Using self-efficacy as a unifying theoretical construct, Wine (1971) postulated that test anxiety was not unidimensional. Since high test anxious persons typically interpret a wide range of situations as evaluative, reacting with cognitive concern and performance deficits, he posited that test anxiety was a cognitive-attentional construct. Assessment of the positive orientations to evaluation with measures of self-efficacy resulted in increased predictive power of the

test anxiety scale. Wine's research resulted a moderate, negative correlation ($r = -.37$) between self-efficacy and test anxiety. Using 350 adolescent subjects, those who scored as high test anxious reacted with self-reports of high emotional reactivity.

Bandura (1977) offered a task-specific conceptual framework of the self-efficacy construct. He proposed that for each study, a unique scale would be developed. Gist (1987) had also adopted this definition of self-efficacy and the means to measure it. In her study, she found decreased predictive power as the self-efficacy measure became more generalized.

In a review of the literature, Gist (1989) offered significant results with situationally specific efficacy scales. She postulated that key efficacy perceptions, when identified, discriminated between high and low performers. However, this held only when vague past performance information was available. She continued research using task-specific efficacy instruments to explore computer efficacy. Significant increases in performance occurred for subjects with high computer efficacy. This particular study looked at self-confidence, prior computer experience and past success in learning situations. The 6-item scale measured efficacy of computer operation over six difficulty levels.

In other related research, Wang and Richarde (1988)

conducted a study to reconcile accepted task-specificity of the efficacy construct with growing evidence that it may be validly measured with more generalized scales. Correspondence between general and specific measures resulted in a bipolar factor, suggesting that each scale assessed opposite aspects of the same construct. They concluded that the generalized measure would most successfully predict performance in situations that were less familiar or perhaps ambiguous to the individual.

In contrast to these and other findings, Riggs (1989) found moderate evidence for the suitability of a more generalized job efficacy scale. His research was conducted in work-related setting. The design of the scale emphasized its applicability across occupational types. Thus the scale was general enough to be used in a variety of work settings, but specific to the workplace. He found personal efficacy correlated with performance ($r = .21$, $p < .01$) and reliably measured a single construct.

CLASSICAL AND MODERN TEST THEORIES

Psychometric justification for use of computers in screening and selection processes must be established in order to encourage widespread use within organizations. Modern test theory, when applied in computer-administered test settings, offers some promising resolutions to problems of subject artifacts.

Entrenchment of modern test theory in place of classical psychometrics is the basis for new, computerized testing models. However, even though arguments against modern test theory are not fully resolved, advantages continue to appear in the literature.

A common characteristic of traditional testing instruments includes the fixed set of items which are administered to all examinees (Weiss, 1982). Scale development is based on comparisons of some internal criteria (i.e., a psychological trait) with emphasis on assessment of item validity. Internal consistency is usually assessed by Cronbach's coefficient alpha (1951). The obtained test score distribution is compared with the test developer's desired test score distribution. This method of test development compounds the difficulty of standardized scoring methods.

Generally, no two tests scores may be compared.

Difficulty and discrimination indices vary according to the distributions of ability in subgroups of examinees. Therefore, traditionally developed tests are not independent of the samples on which their development is based. This sample dependency is readily reflected by changes in the reliability coefficient as a function of the true score variance in the particular sample, despite consistency in the size of the measurement errors. Sample dependency occurs in the classical parameters of item p-values, item-test correlations, and validity coefficients (Hambleton & van der Linden, 1982).

Observed test scores and error of measurement are estimated. For theoretical reasons, one would expect some differences between test-retest scores due to random influences. Construction of exacting parallel measures seems difficult from a practical standpoint. However, parallel testing procedures are necessary in order to avoid problems from repeated administration of the same test.

Guion and Ironson (1983) delineated shortcomings of classical test development as compared with modern test theory, or item response theory (IRT). Sample dependency, a single overall standard error of measurement for all examinees, restriction of comparison due to test score metric, and disregard for the pattern of item responses

all contribute to problems with classical psychometric theory. IRT readily addresses these issues.

Hambleton and van der Linden (1982) presented comparative analysis of classical test theory and IRT. Classical test theory aims at the level of the test whereas IRT aims at the level of the item. IRT establishes a different relationship between the test score and the variable measured by the test. Rather than aggregating item responses as the total score, IRT employs individual item responses, with probabilities of success as a function of the examinee and the item. Individual item responses are used as indicators of ability.

A growing body of evidence shows that test development could be improved if individual item information about item responses were used. Early researchers, such as Lazarsfeld (1950) categorized the estimated ability estimate as a "latent trait."

In monte carlo studies, Lord (1952, 1953) provided strong mathematical evidence for use of latent trait theory in test construction. IRT was later successfully applied to test score equating (Lord, 1975), tailored testing (Lord, 1968; Weiss, 1976), and test design and evaluation (Wright, 1977).

IRT replaces the classical test theory true score estimate with the latent parameter theta (θ), which is not

indexed to the test. The true score scale depends on a specific set of items, but the ability scale does not.

Since IRT modeling begins prior to scoring of the test, quantitative item and ability parameters are used to explain qualitative item responses. IRT analysis uses models of probabilities of $P(\theta)$, as the probability of a specific response at a given ability level θ (Anderson, 1982).

One, two, and three item parameter models are used in IRT. Parameters are determined by item discrimination, item difficulty, and a guessing parameter. Conditional, direct, or marginal maximum likelihood methods are applied to estimate the parameters. Goodness of fit tests such as chi-squared check the appropriateness of the model chosen.

Examinee true score, using classical test theory, will vary across nonparallel measures of the same ability. IRT establishes an ability estimate that would be the same across a sample of items, whether parallel or nonparallel measures. Examinee performance on a measure is transformed to a standardized estimate of ability, θ, as a common scale. Examinees may be compared even when they have not taken the same items. In this respect, IRT is a "test-free" measurement (Crocker & Algina, 1986). The scale of measurement, as an arbitrary unit of measure,

establishes the mean latent trait score as equal to 0 and the standard deviation as equal to 1.

A common assumption of IRT states that a single ability underlies and explains examinee test performance. This single ability is also assumed to be unidimensional. A test is defined as unidimensional if items throughout the test measure a single trait or ability. The latent trait accounts for the statistical dependence among items (Crocker & Algina, 1986).

Local independence is another assumption for the IRT models. This assumption states that the probability of an examinee answering a test item correctly would not be affected by performance on any other item in the test (Hambleton & Swaminathan, 1985).

Latent trait models specify a relationship between observable examinee test performance and an unobservable latent trait. Thus latent traits or abilities are not directly measurable (Hambleton and Swaminathan, 1985).

Another basic assumption of latent trait theory is that examinees' performance on a test can be predicted by defining examinee characteristics or traits. For example, examinee estimated scores on assumed underlying traits are subsequently used to predict test performance (Lord and Novick, 1968).

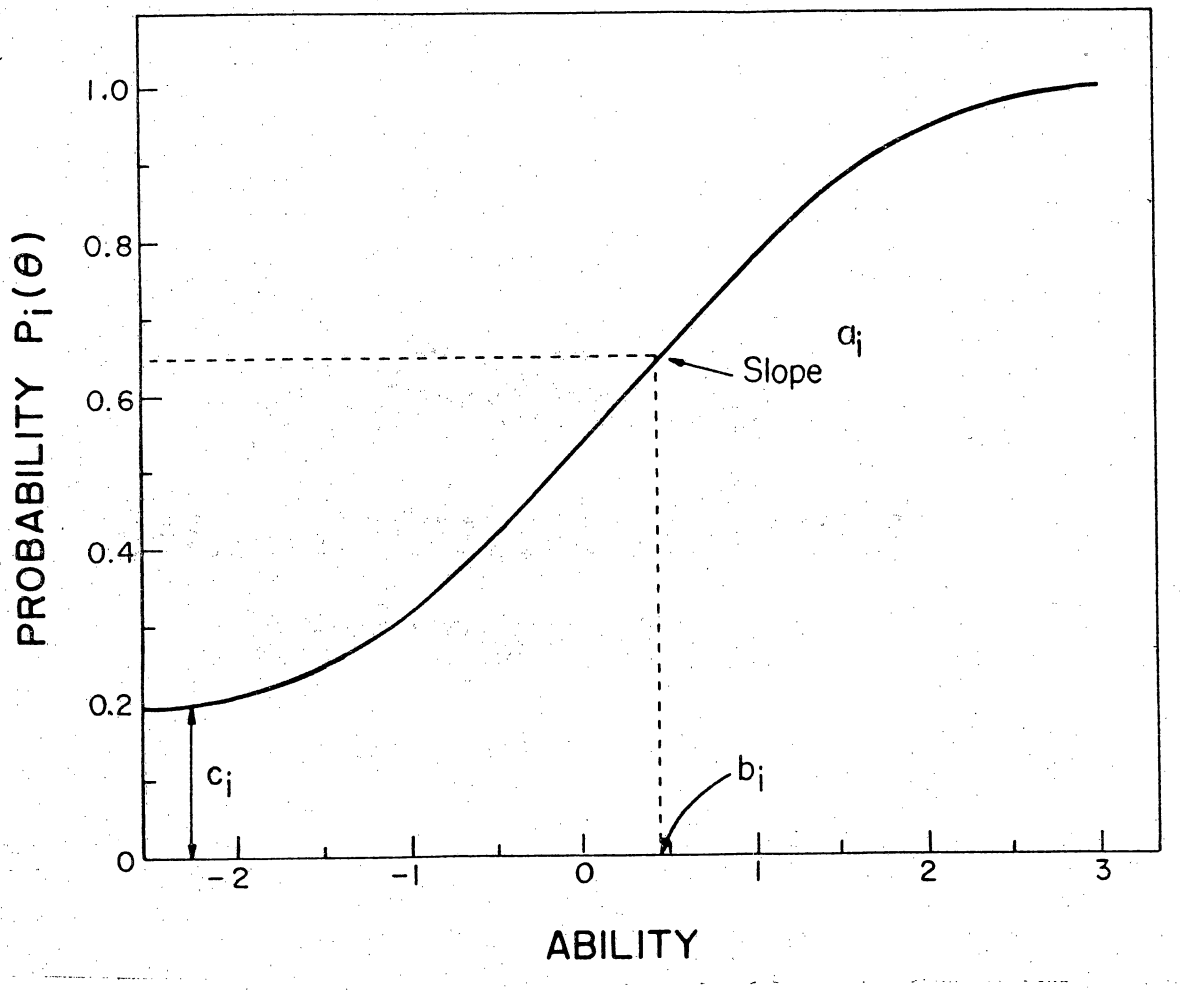
IRT models depend on the number of parameters used. In the Rasch model (Wright, 1977) all items have the same

discrimination parameter. In the two-parameter model, the discrimination and difficulty parameters are indexed. The three-parameter model, or the logistic model (Birnbaum, 1968), includes the guessing parameter. Fischer and Formann (1982) used the three-parameter logistic model since they found it more flexible for items and item formats.

Three, two, and one-parameter IRT models have been reviewed in the literature. Essentially, within each model, item parameters are used to determine the item characteristic curves (ICC; See Figure 1.). The trait scale is placed on the horizontal axis as the level of ability or θ. The probability of a specific response to an item is on the vertical axis. Parameters are: (1) item discrimination, "a_i," (2) item difficulty, "b_i," and (3) the guessing parameter, "c_i."

The slope of the ICC is related to the item discrimination parameter. If the slope of the curve is steep, the probability of a particular response changes rapidly in relation to the changes of the trait level. Items discriminate maximally among examinees at slightly different levels of the trait at the steepest point of the slope. If it may be assumed that all items are of near-equal discrimination, then this parameter may be set at a constant.

Figure 1.



Typical Item Characteristic Curve

Location of the curve along the horizontal axis is a function of the difficulty parameter. This is the point on the latent trait scale at which the slope of the ICC is at a maximum. Increased accuracy of prediction from use of the mathematically-derived ICC occurs, relating the probability of success on an item to the ability measured (Hambleton and Swaminathan, 1985).

The lower left asymptote of the ICC indicates the guessing parameter. This represents the probability of an examinee of low ability correctly answering an item (Guion and Ironson, 1983). If none of the items may be answered correctly by guessing, this parameter may be set to 0, reducing computational time.

The ICC is a nonlinear regression function considered as either a normal ogive or logistic curve, depending on the particular model used. The normal ogive is similar to the cumulative normal distribution. Although used predominantly in early research, it has been replaced recently with the computationally simpler logistic model (Birnbaum, 1968).

IRT has the added advantage of providing equal precision at all levels of the trait continuum being measured. The item pool has highly discriminating items equally represented at the full range of difficulty. Conversely, conventional test construction leads to difficulties such as the "bandwidth-fidelity" dilemma

(Weiss, 1983). "Peaked" conventional tests are most precise at trait levels where the test is peaked, but it does not provide optimal measurement for examinees for whom the test is too easy or too difficult.

IRT therefore takes into account the pattern of item responses. A precise standard error of measurement at each ability level leads to maximum information at that level. In classical psychometrics, a total score based on aggregated item responses is used to compute the overall standard error of measurement. With IRT, the probability of an examinee correctly answering an item depends on the form of the ICC and is therefore independent of the distribution of examinee ability. The probability of a correct response does not depend on how many other examinees are located at the same ability level. This "sample independence" is one of the strongest characteristics of IRT. Compared with classical test theory, it clearly distinguishes itself as the more attractive model.

IRT research has typically been simulations (Urry, 1970; Lord, 1968). Applied research generally has been in educational settings (Olsen, et al., 1986; Pine, 1986; Weiss, 1982).

ADAPTIVE TESTS

In standardized ability tests, when item difficulty is varied, there is the unintended consequence that most examinees must respond to items that are either too easy or difficult to provide information about their ability. Adaptive tests provide an effective solution to this problem. The total number of items required for administration to achieve a specified level of measurement precision is reduced. This reduces boredom, minimizes test fatigue, saves time and money.

Urry (1977) discussed several advantages of the new, computer-adaptive testing technology which included: 1) standardized administration and avoidance of test bias from variations in administration variability, 2) less risk of compromise because tests would no longer be printed, and 3) improved validity and measurement accuracy.

Adaptive tests are a result of IRT models. Lord (1980) found adaptive testing to be a more accurate and equiprecise measurement throughout the range of the trait or ability tested. Item selection procedures are based on an estimation process that computes examinee ability. At each response, the computer chooses an item that would best estimate the examinee's true ability score (θ). This

is based on an initial ability estimate, which is typically derived from prior ability test scores or from item parameters from a normative sample.

Assuming the initial ability estimate is valid, IRT uses an iterative process to select items, optimizing specific criteria. Methods include maximum likelihood estimation (MLE; Lord, 1980), maximum information item selection (Hambleton and Swaminathan, 1985), and Bayesian priors (McBride, 1977).

In a monte carlo study of adaptive testing, Hulin (1983) combined the MLE and maximum information item selection procedures. He found that for examinees in an adaptive test administration with Θ as low as -1.75, Θ was estimated more accurately than with conventional test administration.

Olsen, et al., (1986) and Bock and Mislevy (1982) found that the 3-parameter model generally provided a better fit to the data as compared with the other models. In both studies, computer-adaptive tests produced an ability distribution with a mean closer to 0, as well as smaller variance. Results in both studies were contrasted with those from paper-and-pencil and computer-conventional test administrations.

Test equating studies showed strong possibilities for alternate versions of measures. Huba (1988) developed

alternate versions of the Western Personnel Test (WPT; Gunn and Manson, 1962) which included paper-and-pencil and computer-conventional. Correlations between test forms were high (r = .76) and no significant differences between test groups resulted. He concluded that these alternate versions could be used interchangeably. Thus, paper-and-pencil versions of tests might easily be adapted to administration on computer.

Simulations by Urry (1977) have shown that a model with insufficient parameters led to ineffective adaptive testing. He recommended that test developers carefully review the theoretical implications of their test prior to choice of the model. He cited the an example of the inappropriate use of the Rasch model for tests with multiple-choice items, since this model did not support the fidelity of multiple-choice item response data.

In a study by Olsen, Maynes, Slawson, and Ho (1986) equating and comparison of paper-and-pencil, computer-conventional, and computer-adaptive tests resulted in no significant differences among administrations. Calibrations from 1, 2, and 3-parameter models showed increased test information and reduced standard errors as the number of parameters was increased. These results were similar to those of Bock and Mislevy (1982) in a study of effects of administrations using the Armed Services Vocational Aptitude Battery.

Hulin (1983) pointed out a variety of potential problems with adaptive testing. During the ability estimation process, and "initial item" is chosen. This item is generally administered to all examinees first, no matter what their ability level. It is considered a starting point for the estimation processes. Disagreements have occurred among researchers regarding the importance of this initial item. However, Lord (1977) had provided evidence that choice of the initial item had little or no effect on accuracy of the ability estimates.

Computer-adaptive testing adds another dimension to screening and selection processes. Faced with a computer adaptive instrument, examinees might become fearful that the test is too short to be effective or fear that the computer has malfunctioned. Implications for computer anxiety research must be considered.

Issues of motivation are affected also. Adaptive testing eliminates administration of high-difficulty items to low-ability examinees, and conversely, administration of low-difficulty items to high-ability examinees. It may be predicted that increased efficacy would occur using computer-adaptive testing.

In a review of the literature that focused on anxiety in computer-adaptive conditions, Garrison and Baumgarten (1961) gave entry level college students attitude

questionnaires. Questions were formulated such as: "The amount of time between questions was too fast/too slow...," "Operating the computer was simple/confusing," "while taking the test I was nervous/relaxed." A majority of the subjects responded with positive attitudes, even though nearly half found the use of the computer as more difficult.

Pine (1986) researched the possibility that adaptive testing provided increased motivation. He also assessed the test equivalency between paper-and-pencil methods and computer-adaptive methods. Using a 4-item scale to assess nervousness, he found significant effects of increased anxiety in the adaptive condition. He concluded that the constant matching of examinee ability on an item-by-item basis increased nervousness.

RESEARCH QUESTIONS

The present study used measures of general, test, and computer anxiety, along with self-efficacy and computer-efficacy scales to compare effects among three testing conditions. Equated computer-adaptive, computer-conventional, and paper-and-pencil formats were compared. It was predicted that no significant differences among administration methods would be found.

It was also predicted that the general self-efficacy scale and the computer-efficacy scale would be highly correlated. A negative correlation in the range of .30 was expected between the efficacy and anxiety instruments.

METHOD

Subjects

The test site was a small office at a private, for-profit personnel agency. Individuals who applied for work at a personnel agency were available for testing at an average of 5 to 10 applicants per day. Test batteries were required for placement in occupations such as clerk/typist, secretary, office manager, file clerk and other related clerical jobs. Ss ranged in age from 16 to 73, with both male and female applicants. Only 10 males participated in this study. Demographic statistics are presented in Table 1.

Random assignment to group, or test condition, occurred with each test condition in consecutive order until a total of 50 Ss participated in each condition. Conditions were: computer-adaptive, computer-conventional, and paper-and-pencil.

Ss in all three conditions were briefed, in writing, about the confidential nature of the test results, as well as provided full disclosure of the purpose of the questionnaire. Ss were given the right to terminate participation in the experiment by not turning in the questionnaire data. This addressed the issue of informed consent. Out of 150 total Ss, ten were eliminated due to

lack of complete test data. All but four Ss turned in the questionnaire.

Full disclosure of the purpose of the test took place immediately following the test, in writing. Treatment of participants was in accordance with the ethical standards as presented by the American Psychological Association.

Instruments

This study used measures of clerical aptitude as the test content. Clerical aptitude tests are measures of a specific aptitude or ability which emphasize perceptual speed and accuracy. Anastasi (1988) defined aptitude as a cumulative influence of multiple experiences. Clerical skills typically demand a large portion of time spent on tasks that require speed and accuracy to perceive details.

Tests of alphabetizing skills are considered job sample tests as opposed to aptitude measures. The "file-drawer" aptitude test was therefore a job sample test. The "name/number comparison" test was considered a clerical aptitude test, with scores depending on speed and accuracy. Anastasi (1988) found a marked and consistent difference in favor of women for such skills.

Clerical aptitude, along with other aptitude, achievement and ability measures, have been found to be valid predictors of performance on the job and in training for all jobs in all settings (Schmidt and Hunter, 1981).

Through validity generalization, it was therefore established that cognitive ability tests are equally valid for both minority and majority applicants. Cognitive test validities are generalizable with confidence across organizations and settings.

Additional research by Schmidt and Hunter (1981) resulting from data of 370,000 clerical workers provided consistent validity measures across five different task-defined clerical job families.

A review of available clerical aptitude batteries, such as the Clerical Abilities Battery and the Minnesota Clerical Test provided guidance for item construction. After the clerical tests were constructed, they were administered to 200 subjects in Study 1[1].

Three equivalent versions of the two tests were developed (i.e., computer-adaptive (CA), computer-

---

[1]Item parameters were calculated based on responses by 200 Ss scores on the "file-drawer" and "name/number comparison" tests. Ss were recruited from undergraduate and graduate students at a Southern California University, ranging in age from 18 to 55, including males and females. Tests were administered in a classroom setting using portable computers over a period of six weeks. Both tests had 108 items each. Ss entered responses and data were collected on computer diskettes, analyzed, and used to produce item parameters. Only ten items of the "file-drawer" test were deleted due to no variance. Item parameters were based on the 3-parameter logistic model (Birnbaum, 1968). An adaptive version of the tests was generated using a computer program designed for this purpose (MicroCAT Testing System, version 3.0; Microcat Assessment Systems Corporation, 1989).

conventional (CC), and paper-and-pencil (PnP).

In the computer-administered conditions, Ss used a portable computer to enter responses. The CA condition matched item difficulty to previous examinee responses and therefore all subjects were not administered the same items. The CC condition presented all items in the same order as was presented in the PnP condition. Ss in the CC condition entered responses on the computer keyboard. All Ss in the computer conditions were briefed verbally by agency staff about how to enter the responses and were prompted on keyboard familiarity.

The PnP condition consisted of all test items administered using paper and pencil. Ss were given the test and sat in the testing area of the office. This method included written instructions for the test, along with practice items.

A self-report questionnaire was developed to assess anxiety and efficacy (See Appendix A). General anxiety was measured by the State Trait Anxiety Index (STAI; Spielberger, Luchene, and Goshen, 1970). Only State anxiety was assessed using a 20-item format. Trait anxiety was not assessed due to the nature of the construct as a stable personality characteristic.

Reliability estimates for the State anxiety scale were estimated to be in the range of .91 to .94 (Spielberger, 1966). Significant correlations between

State anxiety scale and Trait anxiety scale were found in the range of .70 to .75 in working adults (Spielberger, 1966). Stronger correlations were found in social-evaluative situations. Convergent and divergent validity reported by Spielberger (1980) positions the State anxiety scale as a solid measure.

Test anxiety was measured using the Test Anxiety Index (TAI; Spielberger, 1978). No changes or deviations from the standard 20-item scale occurred. Reliability of the TAI was reported in the range of .94 to .95 (Spielberger, 1978). The TAI was found to correlate .56 with the Test Anxiety Scale (TAS; Sarason, 1978), confirming its use as a situation-specific measure of anxiety proneness during tests.

Computer anxiety was assessed with the appropriate factor from the Computer Anxiety Rating Scale (CARS; Rosen, et al., 1988), a 19-item scale. Reliability of the scale ranged from .93 to .95 (Rosen, et al., 1988). Rosen cautioned on use of only a single factor from his research without verification through replication. However, use of the scale in the present study was perceived as acceptable due to the high face validity of the items. Additionally, Rosen reported that this factor accounted for 40.3% of the total variance explained in computer anxiety, as measured by his multi-dimensional scale.

Efficacy was assessed with both task-specific and general scales. Gist (1989) presented a specific computer-efficacy scale consisting of six items. Her research focused on efficacy in a training setting. The items were rated on a scale of 1 to 10, with a "can do/cannot do" response. For the purposes of this study, the question stems were used but with a more standard, 1 to 6 Likert format. The effects of this change on the psychometric properties of this scale could not be assessed, but were believed to be negligible.

The general self-efficacy scale (Riggs, 1989) was also administered based on its strong face validity. This 12-item scale was expected to contribute additional information about mean group differences on anxiety measures. The research conducted by Riggs included measures of outcome expectancy. The test-retest reliability was .80.

The combination of all scales produced a questionnaire of 75 items, which was presented in a paper-and-pencil format. Ss circled the answer that corresponded to their response. Questionnaires were placed in a sealed envelope to protect confidentiality.

Procedure

Ss entered the office of the personnel agency and requested an application for employment. They were handed

a clipboared and seated in the testing area. Agency staff reviewed the completed application for demographic data such as age, years clerical experience, total years education, years computer experience, computer ownership, and sex.

All Ss were tested. Each S was assigned a random identification number. Ss sat at either the computer terminal or in the testing area, depending on the assigned test condition. Staff of the personnel agency were briefed on the purpose of the experiment, the use of the computer, and how to collect test data. On site data collection took place over a period of approximately five months.

Once the testing was completed, all Ss were given the questionnaire, which included an introductory statement underscoring the confidentiality of the responses and that responses would not influence a decision for hire. This disclosure was repeated throughout. Approximately 20 extra minutes was required to fill out the questionnaire.

After completion of the questionnaire, Ss were presented with a written statement of disclosure regarding the purpose of the experiment.

The experiment had 3 conditions: 1) computer-adaptive, 2) computer-conventional, 3) paper-and-pencil. Testing was followed by administration of the questionnaire with five scales consisting of: 1) State

anxiety (STAI), 2) test anxiety (TAI), 3) computer anxiety factor from the computer anxiety rating scale (CARS), 4) general self-efficacy , and 5) computer-efficacy.

Results from the clerical tests were kept separate from the questionnaire data collection procedure. In order to preserve the experimental conditions, informed consent took place after each S had completed both the clerical tests and the questionnaire. Actual scores for placement were scored and ranked according to percentile norms by agency staff. Agency employees had no access to responses on the questionnaire. Ss were asked to place the questionnaire in to an envelope and seal it. Only a randomly-assigned identification number appeared on the outside.

Applicants were then asked to interview for job placement. Test scores from the clerical batteries were used for screening/selection purposes.

RESULTS

A one-way ANOVA was conducted to compare the mean group differences among the three conditions using measures of state, test, and computer anxiety and efficacy scales, as well as on demographic variables. No significant differences among experimental conditions were found. For each ANOVA, the standard error of measurement for the dependent variable of interest was used as an effect size estimate, allowing for estimation of the statistical power. Power was uniformly high, ranging from .70 to .90.

In addition to conducting statistical tests for significance of group mean differences using one-way ANOVAs, additional light was shed on the magnitude of the group differences by the calculation of credible intervals (Hays and Winkler, 1971), also referred to as High Density Regions (HDR; Schmitt, 1969).

Credible intervals provide a probabilistic means by which to weight interpretation of the outcome of no significant differences. As a scale of credibility, they are presented in a familiar metric. Like 95% confidence intervals, credible intervals may be interpreted as a 95% chance the true mean difference is in the interval. Ends of the interval are compared, relative to 0, to the

standard error of measurement for the dependent variable of interest. If an interval includes 0, the probability is low that there is a large difference among group means and a Type II error has occurred. If the interval does not include 0, then there is not unequivocal support for the claim of no differences.

The ANOVA on the State anxiety from the STAI resulted in no significant differences. Results are presented in Table 2.

Table 2

STAI

Source	DF	SS	MS	F Ratio	F Probability
Between	2	59.81	29.91	.33	.72
Within	124	11246.41	90.70		
Total	126	11306.22			

A 95% credible interval was calculated for STAI comparing CA to CC (HDR = -2.0350 to 1.5150). This interval included 0. The magnitude of the differences between the means was compared to the standard error of measurement for STAI (SEM = 2.81). Coupled with the credible interval estimate, this comparison suggests that

a conclusion of no substantive differences among groups on the STAI remains tenable. This analytic approach was continued for each pairwise comparison of the experimental conditions for each dependent variable.

The credible interval for the difference between CC and PnP ranged from -1.0063 to 3.5357, again relative to the SEM of 2.81.

The credible interval for the difference between CA and PnP also substantiated the result of no difference (HDR = -.7207 to 2.8728; SEM = 2.81.

The ANOVA conducted on TAI resulted in no significant differences. Results are presented in Table 3.

Table 3

TAI

Source	DF	SS	MS	F Ratio	F Probability
Between	2	226.55	113.27	.85	.43
Within	124	16576.68	133.68		
Total	126	16803.23			

Calculated 95% credible intervals for the TAI resulted in a difference between CA versus CC that ranged from -1.13 to 5.06; CA versus PnP ranged from .7015 to 5.7585; and CC versus PnP ranged from -.3540 to 2.840.

These intervals should be interpreted relative to the standard error of measurement which was 3.0634. Only the CA versus PnP interval does not include 0.

Similar results were achieved with the ANOVA on the CARS factor of computer anxiety. No significant differences were found. These results are reported in Table 4.

Table 4

Computer Anxiety

Source	DF	SS	MS	F Ratio	F Probability
Between	2	228.90	114.45	.63	.53
Within	122	22090.29	181.07		
Total	126	22319.20			

Computation of the 95% credible intervals for the CARS factor of computer anxiety resulted in the difference between CA versus CC ranging from -2.83 to 3.19 and CC versus PnP ranging from -.07 to 5.54, with both intervals including 0. These should be compared to the SEM = 3.32, with the conclusion of no substantive differences.

The conditions of CA versus Pnp resulted in the difference ranging from .08 to 5.75. This did not include

0, but the lower end of the interval was close to 0 when contrasted to the SEM = 3.32. Marginal non-significance was found in the one-way ANOVA for the self-efficacy scale ($F(2,122) = 2.66$, $p = > .05$). The CA condition had a higher mean score ($M = 56.14$).

The results of the ANOVA for self-efficacy are presented in Table 5.

Table 5

Self-Efficacy

Source	DF	SS	MS	F Ratio	F Probability
Between	2	417.45	208.72	2.66	.07
Within	122	9559.07	78.35		
Total	124	9976.51			

Calculation of the 95% credible interval for CA versus CC ranged from 1.22 to 5.06, with SEM = 3.02. The interval for CA versus PnP was 2.24 to 6.13. Neither of these intervals included 0. With other dependent variable group comparisons, the credible interval either included 0 or had a lower bound that was close in absolute magnitude and close to 0 relative to the SEM. However, with the self-efficacy measure, mean differences with experimental conditions produced credible intervals that did not

include 0. The lower bounds of the interval were also not close to 0 in absolute terms and different from 0 by a substantial portion of the SEM.

The credible interval for CC versus PnP was -.72 to 2.52m which included 0.

Results of the ANOVA with findings of no differences among groups for the computer efficacy variable are presented in Table 6.

Table 6

Computer Efficacy

Source	DF	SS	MS	F Ratio	F Probability
Between	2	30.49	15.25	.58	.56
Within	132	3445.51	26.10		
Total	134	3476.00			

Credible intervals for computer efficacy were not calculated because instrument was not used in any manner similar to that reported in the literature. No prior estimate of variance from the scale using the 1 to 6 Likert format was available from which to calculate the intervals.

Analysis of the relationships among the five scales

yielded moderate correlations between general and test anxiety ($r = .51$, $p < .001$). Computer anxiety was not significantly related to general anxiety ($r = .09$, $p > .05$).

Negative correlations between the anxiety scales and measures of efficacy occurred, as was supported by the literature (Wine, 1971). Additionally, general self-efficacy correlated with computer anxiety ($r = -.28$, $p < .001$), as was predicted.

A moderate correlation occurred between general and computer efficacy measures ($r = -.37$, $p < .001$). General self-efficacy accounted for 24% of the variance in test anxiety. Also, 26% of the variance in general anxiety could be accounted for by test anxiety. Results of the correlation matrix are presented in Table 7.

Reliability coefficients (using Cronbach's Alpha), reported in Table 1, for all scales were acceptably high. The lowest reliability occurred with the STAI ($r = .84$).

Table 7

Correlation Matrix

	General Anxiety	Test Anxiety	Self Efficacy	Computer Anxiety	Computer Efficacy
General Anxiety	1.000	.52**	-.50**	.09-	.26*
Test Anxiety	.52**	1.000	-.37**	.24*	-.25*
Self-Efficacy	-.50**	-.37**	1.000	-.30*	.38**
Computer Anxiety	.09	.24*	-.28*	1.000	-.33**
Computer Efficacy	-.26*	-.25*	.37**	-.33**	1.000

N of CASES: 99 1-Tailed Significance: * .01
** .001

Table 1

Overall and Group Dependent Variable Means, Standard Deviations, and Alpha Measures

	N	Mean	Standard Deviation	Cronbach's Alpha
STAI		36.71	9.48	.84
CA	42	36.07	1.20	
CC	39	36.33	8.55	
PnP	46	37.61	11.54	
TAI		36.12	11.55	.97
CA	40	37.93	12.11	
CC	42	35.95	10.75	
PnP	45	34.67	11.80	
CARS		31.68	13.42	.96
CA	40	32.75	13.90	
CC	42	32.57	12.62	
PnP	43	29.81	12.62	
SE		53.65	8.97	.87
CA	42	56.14	9.82	
CC	41	52.93	8.12	
PnP	42	51.86	8.51	
CE		30.00	5.09	.91
CA	44	30.41	5.45	
CC	45	29.33	5.05	
PnP	46	30.26	4.83	

Analysis of the demographic data resulted in a significant difference among the experimental groups in full-time work experience ($F(2,137) = 8.51$, $p < .001$). The PnP condition ($M = 8.06$) had the highest full-time years of work experience as compared to the CA condition ($M = 3.11$) and the CC condition ($M = 4.54$). No other significant differences between groups were found on other demographic variables.

To summarize, statistical analysis consisted of oneway ANOVAs on variables to compare group means. No differences among conditions were found on measures of general, test, and computer anxiety, along with general self-efficacy and computer efficacy. The credible intervals suggested that there were no substantive differences.

DISCUSSION

This study was designed to provide evidence that computerized and computer-adaptive tests would not adversely affect anxiety levels or efficacy when assessing clerical skills. The results of this study clearly suggest that this is the case.

These results accentuate findings by Schmidt, Hunter, McKenzie, and Muldrow (1979) regarding the economic utility of valid selection procedures. Their review of the literature emphasized increased productivity of workers tested with innovative testing technologies such as computers. They found that innovative selection procedures reduced test time and cost of administration, allowing for increases in the total number of applicants screened.

Hambleton, et al., (1978) pointed out the advantages of adaptive tests and the latent trait theoretical approach to resolve mental measurement problems. This study contributed to the growing body of evidence that applied latent trait theory and use of computers in testing situations actually help eliminate some of the long-standing measurement dilemmas such as length of test, fatigue, practice effects, and others while not increasing anxiety.

Review of the performance of the separate dependent measures produced interesting outcomes. The STAI performed equally well in this study as it did in research conducted by Spielberger (1978). Mean scores were comparable. However, it should be pointed out that Spielberger's data was primarily gathered in educational and clinical settings. He also focused his research on issues of validity, as did most other studies (Spielberger and Sarason, 1961). This study, on the other hand, looked only at the effects among methods of test administrations.

Hedl and O'Neil (1977) found reduced State anxiety under conditions of computer-based learning. Their study presented the computer testing environment as less anxiety-provoking overall. While the current study did not replicate the findings of less anxiety, adverse effects were not increased.

Analysis of data from the present study found lower mean scores of the TAI in all three conditions as compared to research and normative data presented by Sarason (1978). The lowest mean value occurred in the paper-and-pencil condition. Ss expressed less worry about test results and reduced cognitive concern. Sarason's research was conducted using students who were less likely to have even moderate years of work experience. In the present study, Ss had not only work experience, but also clerical

experience. Thus the test itself was not new material, and was not perceived as threatening.

Perceptions of an evaluative condition were minimized for all conditions by use of simple test instructions and a set of practice items preceding each clerical test. Non-achievement orienting instructions were found to affect the levels of general and test anxiety in a study by Sarason (1978). A reduction in the evaluative threat resulted in less general and test anxiety. Both this study and Sarason's had similar results. However, as pointed out earlier, Sarason conducted the study in educational settings. The present study therefore extends these findings into the workplace.

In an analysis of correlates of computer anxiety, Morrow, et al., (1986) found attitudinal or personality variables did not explain as much variance as prior experiences. However, researchers defined experience as prior, hands-on use of computers. Their research pointed out the possibility that computer anxiety may be a modifiable condition.

Spielberger (1966) had also found test stimuli contributing to the level of test anxiety. Thus, reduction of the evaluative threat, such as changes in the test instructions would help decrease anxiety levels in examinees. This study used not only non-threatening instructions, but also provided practice items for the

clerical tests.

The CARS factor of computer anxiety produced mixed results. A non-significant correlation was found with the TAI. The correlation with STAI was only modest. These results accounted for less than 1% of the variance found in the STAI, and approximately 5% of the variance in the TAI. Rosen (1988) had cautioned against the use of any single factor of the CARS, stating that interpretation should be treated prudently. The observed, mixed results support this caution. However, given the applied setting in which the CARS factor was used, it still yielded no significant differences among the group means and high reliability (alpha = .96).

Items found in the computer anxiety scale had high face validity and therefore performed well in the applied workplace setting. Nonetheless, Ss in the PnP condition did not use a computer. Many Ss failed to respond to these items. For this reason, the only genuinely accurate comparison could be made between the CA and CC conditions.

Mean scores on computer anxiety were found to be lower overall in the present study (CA; M = 32.75, CC; M = 32.57, PnP; M =29.81). This may be due to response sets. Social desirability was stronger as a result of the screening/ selection process. Ss wanted to appear less anxious about the use of computers, reflecting an

acceptance of the inevitable appearance of computers in the workplace.

Ss in the PnP condition had the lowest scores in computer anxiety, since they did not use a computer for testing. However, no significant differences occurred among conditions. Applicants may have wanted to appear more likely to be easily trained, and more willing to work with computers on the job. It should be pointed out that items on the computer anxiety scale were worded so that all Ss could respond.

Efficacy measures were included to assess other possible reactions to the testing process. Ss in the CA condition had the highest scores for general job self-efficacy (M = 56.14) as compared with the CC condition (M = 52.93) and the PnP condition (M = 51.86). Two studies (Brod, 1982; Hill, et al., 1987) also found that increased self-efficacy beliefs contributed to the ability to adapt to innovative technologies such as the use of computers. Despite the results of no significant differences among group means, the higher mean for the CA condition suggests some relevance to the Brod (1982) and the Hill, et al.(1987) studies. Continued research in this direction may uncover a trend toward more efficacious beliefs in computer conditions.

Ss in the CA condition had the highest levels of self-efficacy. Ss in this condition may have considered

the adaptive test as a "cutting edge" experience in the screening/selection process. Since computers continue to penetrate the workplace, applicants may have viewed their experience as an encounter with the office tool of the future. The fact that they simply were able to use it may have affected their sense of self-confidence and ability to perform well.

Gist (1989) found that task-specific efficacy measure most useful in determining prior task mastery. Since applicants in the present study had a substantial amount of prior work experience, it follows that their self-reported measures of efficacy would indicate a stronger sense of capability to perform a task. Results of this study confirm these findings.

Comparisons of the CARS computer anxiety factor with the efficacy measures yielded moderate negative correlations ($r = -.28$ $p < .01$ for general job self-efficacy; $r = -.33$, $p < .01$ for computer efficacy). If high anxious examinees experienced performance decrements, and efficacy measures have been found to increase the predictive validity of anxiety scales, this relation is reestablished in this study (cf., Wine, 1971).

Furthermore, the task-specific computer efficacy measure may be useful as a predictor of high performance. The study by Hill, Smith, and Mann (1987) found high self-

efficacy as an indicator of willingness to use a computer, and helped distinguish those who were computer illiterates.

A negative correlation was found between general job self-efficacy and the STAI (r = -.49). In a study with parallel results, Wine (1971) found a -.36 correlation between the two scales. The general efficacy items had high face validity for job applicants. Questions were work related and might have easily been considered task-specific in the applied setting.

The higher negative correlation in the present study contradicts the findings by Gist (1989). Her study found decreased predictive utility of more general efficacy measures. However, her study focused on training aspects of computer use.

This study presented a cumulative corroboration of the hypothesis of no difference among testing methods. The results seem generalizable based on the applied setting of the study. Ss in the sample used were probably typical of employment agency applicants in general.

The outcome of this study has favorable utility for continued research. Overall, the computer-adaptive method does not present significant obstacles to test-taking strategies as delineated by Burke and Normand (1987), nor does it significantly affect levels of anxiety. As computer-administered testing continues to expand into all

realms, the results reported here suggest examinees will not experience higher levels of anxiety.

Examinees tested by computer-adaptive instruments do not experience increased levels of anxiety even though they are faced with a format very different from the paper-and-pencil method. They cannot review test questions in advance, nor can they necessarily repeat questions, especially with timed tests. Nonetheless, this study showed evidence that this new experience did not significantly affect anxiety levels.

Applicant self-efficacy was found to be associated with anxiety levels. With a need for ever more accurate screening and selection procedures, employers may be faced with the need to assess trainability and worker confidence in relation to computers. Computer-adaptive technologies, along with conventional computer use, do not present any additional problems to either employer or prospective employee.

This research has contributed to the understanding of the theories of applied modern test theory. It has additionally provided a basis for additional studies in applied settings. Further research on this might include and expanded subject pool to cross-validate findings. The issue of computer anxiety has yet to be fully explored, with an adequate scale to measure the construct. Other

extensions of this research might question whether these findings are generalizable to other abilities besides clerical skills, such as spatial, mental, psychomotor, mechanical, and others.

This study provides a strong foothold for future test developers, opening the door to new testing technology. Since no increases in anxiety were found in the computer-adaptive administration, statisticians and test constructors may begin to readily access item response theory for a more accurate, efficient, and quicker method of testing.

APPENDIX "A"

QUESTIONNAIRE

PART 1.

A number of statements which people have used to describe themselves are given below. Read each statement and circle the number that corresponds to your present feelings. There are no right or wrong answers. Do not spend too much time on any one statement.

1 = not at all
2 = moderately so
3 = somewhat
4 = very much so

1.	I feel calm.....	1	2	3	4
2.	I feel secure.....	1	2	3	4
3.	I am tense.....	1	2	3	4
4.	I feel strained.....	1	2	3	4
5.	I feel at ease.....	1	2	3	4
6.	I feel upset.....	1	2	3	4
7.	I am presently worrying over possible misfortunes.....	1	2	3	4
8.	I feel satisfied.....	1	2	3	4
9.	I feel frightened.....	1	2	3	4
10.	I feel comfort.....	1	2	3	4
11.	I feel self-confident.....	1	2	3	4
12.	I feel nervous.....	1	2	3	4
13.	I am jittery.....	1	2	3	4
14.	I feel indecisive.....	1	2	3	4
15.	I am relaxed.....	1	2	3	4
16.	I feel content.....	1	2	3	4
17.	I am worried.....	1	2	3	4
18.	I feel confused.....	1	2	3	4
19.	I feel steady.....	1	2	3	4
20.	I feel pleasant.....	1	2	3	4

PART 2.

1 = almost never
2 = sometimes
3 = often
4 = almost always

21.	I feel confident and relaxed during tests.....	1	2	3	4
22.	I have an upset, uneasy feeling during important tests.....	1	2	3	4
23.	Thinking about the final score on tests interferes with my ability to take the test.....	1	2	3	4
24.	I freeze up on important tests.....	1	2	3	4
25.	During tests, I think about whether I'll succeed.....	1	2	3	4
26.	I become confused when working on tests..	1	2	3	4
27.	Thoughts of doing poorly interfere with my concentration while taking a test.....	1	2	3	4
28.	I feel jittery during tests.....	1	2	3	4
29.	I feel anxious during tests, even when I'm well prepared.....	1	2	3	4
30.	I feel uneasy before getting the results of my test.....	1	2	3	4
31.	I feel tense during tests.....	1	2	3	4
32.	I wish tests did not bother me so much.....	1	2	3	4
33.	I get so tense that my stomach gets upset during tests.....	1	2	3	4
34.	I defeat myself on tests.....	1	2	3	4
35.	I feel panicky during tests.....	1	2	3	4
36.	I worry before important tests.....	1	2	3	4
37.	I am thinking of failing during tests.....	1	2	3	4
38.	My heart beats fast during tests.....	1	2	3	4
39.	I continue to worry after the test is over.....	1	2	3	4
40.	I get nervous and forget facts during a test.....	1	2	3	4

PART 3.

Think about your ability to do the tasks required by your job. If you are not currently employed, think about the job you have applied for or are seeking. When answering the following questions, answer in reference to your own personal work skills and ability to do your job.

REMEMBER--YOUR ANSWERS WILL HAVE NO INFLUENCE OR EFFECT ON HIRING DECISIONS. IF YOU HAVE ANY QUESTIONS ABOUT THIS, PLEASE ASK THE TEST ADMINISTRATOR NOW.

1 = strongly disagree
2 = disagree
3 = disagree somewhat
4 = agree somewhat
5 = agree
6 = strongly agree

41.	Few people in my line of work can do my job better than I can.....	1	2	3	4	6
42.	I have the confidence in my ability to do my job.....	1	2	3	4	6
43.	I enjoy doing my job.....	1	2	3	4	6
44.	There are some tasks required by my job that I cannot do well.....	1	2	3	4	6
45.	When my performance is poor, it is due to my lack of ability.....	1	2	3	4	6
46.	I doubt my ability to do my job.....	1	2	3	4	6
47.	I have all the skills needed to perform my job very well.....	1	2	3	4	6
48.	Most people in my line of work can do this job better than I can.....	1	2	3	4	6
49.	I am an expert at my job.....	1	2	3	4	6
50.	My future in this job is limited because of my lack of skills.....	1	2	3	4	6
51.	I am very proud of my job skills and abilities.....	1	2	3	4	6
52.	I feel threatened when others watch me work.....	1	2	3	4	6

PART 4.

HOW NERVOUS OR ANXIOUS WOULD YOU FEEL IF:

1 = not at all
2 = a little
3 = a fair amount
4 = much
5 = very much

51.	You took a class in computer language (e.g., Basic, Cobol, etc.).....	1	2	3	4	5
52.	You applied for a job that required some training in computers.....	1	2	3	4	5
53.	You reviewed and looked at a computer printout.....	1	2	3	4	5
54.	You visited a computer center.....	1	2	3	4	5
55.	You used a computer program to write a report.....	1	2	3	4	5
56.	You got an "error message" from the computer.....	1	2	3	4	5
57.	You were asked to learn to write computer programs.....	1	2	3	4	5
58.	You dealt with billing errors that were caused by a computer.....	1	2	3	4	5
59.	You learned to use the computer keyboard.....	1	2	3	4	5
60.	You used a calculator that has statistical functions such as means, standard deviations, and correlations.....	1	2	3	4	5
61.	You used a pre-packaged computer program to balance your checkbook.....	1	2	3	4	5
62.	You attended a workshop on uses of computers.....	1	2	3	4	5
63.	You erased or deleted material from a computer file.....	1	2	3	4	5
64.	You thought about purchasing pre-packaged programs for a computer (software).....	1	2	3	4	5

PART 4. (Continued)

1 = not at all
2 = a little
3 = a fair amount
4 = much
5 = very much

65.	You took a class about the uses in computers.....	1	2	3	4	5
66.	You learned computer technology.....	1	2	3	4	5
67.	You thought about buying the hardware to go along with a personal computer (e.g., disk drive, modem).....	1	2	3	4	5
68.	You thought about having to take a class that required limited use of a computer.....	1	2	3	4	5
69.	You used a computer to do statistical computations.....	1	2	3	4	5

PART 5:

I WOULD BE CAPABLE OF OPERATING A MICROCOMPUTER:

1 = strongly disagree
2 = disagree
3 = disgree somewhat
4 = agree somewhat
5 = agree
6 = strongly agree

70.	When I am provided with written instructional material.....	1	2	3	4	56
71.	When the computer gives me instructions at each step, and informs me when I have completed a step successfully.....	1	2	3	4	56
72.	When I am able to listen to someone giving instructions, who pauses as I complete each step.....	1	2	3	4	56
73.	When I am able to watch someone going through the steps before I try the procedures myself.....	1	2	3	4	56
74.	When there is an instructor to watch me as I complete each step, and give me feedback about the correctness of my actions.....	1	2	3	4	56
75.	When there is an instructor to guide me by Telling me each step as I proceed, and explaining the steps and any errors I make.....	1	2	3	4	56

REFERENCES

Anastasi, Anne. Psychological Testing. New York: Macmillan Publishing Company, 1988.

Anderson, Erling B. (1982). Latent trait models and ability parameter estimation, Applied Psychological Measurement, 6, 4, Fall, 445-461.

Assessment Systems Corporation. (1989). User's Manual for the MicroCAT testing system (3rd ed.) St. Paul, MN: Author.

Bandura, Albert. (1977). Self-efficacy: Toward a unifying theory of behavioral change, Psychological Review, 84, (2), 191-215.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M. R. Novick, Statistical theories of mental test scores. Reading M.A.: Addison-Wesley, 1968.

Bock, R. Darrell and Mislevy, Robert J., Adaptive EAP estimation of ability in a microcomputer environment, Applied Psychological Measurement, 6, No. 4, Fall 1982, pp. 431-444.

Bracker, J.S. and Pearson, J.N. (1986). Worker obsolecence: the dilemma of the 80s, Personnel Administrator, 31, 109-116.

Brod, Craig. (1982). Managing technostress: Optimizing the use of computer technology, Personnel Journal, October, 753-757.

Burke, Michael J. and Normand, Jacques. (1987). Computerized psychological testing: Overview and critique, Professional Psychology, Research, and Practice, 18, 1, 42-51.

Crocker and Algina, Introduction to Classical and Modern Test Theory, New York: Holt, Rinehart, and Winston, 1986.

Culler, Ralph E. and Holahan, Charles J. (1980). Test anxiety and academic performance: The effects of study-related behaviors, Journal of Educational Psychology, 72, (1), 16-20.

Fischer, Gerhard H. and Formann, Anton K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters, Applied Psychological Measurement, 6, 4, 1982, pp. 397-416.

Fleishman, Edwin A. (1988). Some new frontiers in personnel selection research, Personnel Psychology, 41, 679-697.

Garrison, Wayne M. and Baumgarten, Bruce S. (1986). An application of computer adaptive testing with communication handicapped examinees, Educational and Psychological Measurement, 46, 23-35.

Gist, Marilyn E. (1987) Self-efficacy: Implications for organizational behavior and human resource management, Academy of Management Review, 12, 3, 472-485.

Gist, Marilyn, Schwoerer, C., and Benson, Rosen. (1989). Effects of alternative training methods on self-efficacy and performance in computer software training, Journal of Applied Psychology, 74, 6, 884-891.

Gist, Marilyn, Benson, R., and Schwoerer, C. (1988). The influence of training method and trainee on the acquisition of computer skills, Personnel Psychology, 41, 255-265.

Guion, Robert M. and Ironson, Gail H., Latent trait theory for organization research, Organization Behavior and Human Performance, 31, 1983, pp. 54-87.

Hambleton, Ronald K., and Swaminathan, Hariharan. Item Response Theory: Principles and Applications, Kluwer: Nijhoff Publishing, 1985.

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

Hambleton, Ronald K. and Cook, Linda L. Robustness of Item response models and effects of test length and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), New Horizons in Testing. New York: Academic Press, 1983.

Hays, W.L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.

Hedl, J. and O'Neil, H. Reduction of state anxiety via instructional design in computer-based learning environments. In J. Sieber, H. O'Neil, and S. Tobias (Eds.), Anxiety, Learning, and Instruction. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

Hill, Thomas, Smith, Nancy D., and Mann, Millard F. (1987). Role of efficacy expectations in predicting the decision to use advanced technologies: The case of computers, Journal of Applied Psychology, 72, 2, 307-313.

Holroyd, Kenneth A., Westbrook, T., Wolf, M., and Badhorn, E. (1978). Performance, Cognition, and Physiological Responding in Test Anxiety, Journal of Abnormal Psychology, 87, 4, 442-451.

Huba, George J. (1988). Comparability of traditional and computer western personnel test (wpt) versions, Educational and Psychological Measurement, 48, 957-959.

Hulin, C.L., Drasgow, F. and Parsons, C.K. Item Response Theory: Applications to Psychological Measurement, Homewood, Ill: Dow Jones-Irwin, 1983.

Lazarsfeld, P.F. The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al., Measurement and prediction. Princeton: Princeton University Press, 1950.

Liebert, R.M. and Morris, L.W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data, Psychological Reports, 20, 975-978.

London, Manual and Bray, Douglas W. (1980). Ethical issues in testing and evaluation for personnel decisions, American Psychologist, 35, (10), 890-900.

Lord, Frederick M., and Novick, M.R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Lord, Frederic M. (1975). The 'ability' scale in item characteristic curve theory, Psychometrika, 44, 205-217.

Lord, Frederic M. (1977). Practical applications of item characteristic curve theory, Journal of Educational Measurement, 14, 117-138.

Lord, Frederic M. Applications to Item Response Theory to Practical Test Problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, Frederic M. (1982). Standard error of an equating by item response theory, Applied Psychological Measurement, 6, No. 4, Fall, pp. 463-472.

McBride, J.R. (1977). Some properties of a Bayesian adaptive testing strategy, Applied Psychological Measurement, 1, 121-140.

Miura, Irene T. (1987). The relationship of computer self-efficacy expectations to computer interest and course enrollment in college, Sex Roles, 16, 5/6, 303-311.

Morrow, Paul C., Prell, Eric R., and McElroy, James C. (1986). Attitudinal and behavioral correlates of computer anxiety, Psychological Reports, 59, 1199-1204.

Olsen, James B., Maynes, Dennis D., Slawson, Dean, and Ho, Kevin (1986). Comparison and equating of paper-administered, computer-administered and computer-adaptive tests of achievement. A paper presented at the American Educational Research Association meeting, San Francisco.

O'Neil, Jr. Harold, F. and Richardson, Frank C. Test Anxiety reduction and computer-based learning environments. In I. G. Sarason (Ed.) Test Anxiety: Theory, Research,and Applications. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

Paul, L. (1982). Research on cyberphiliacs, cyberphobiacs reveals 30% of workers fear computers, Computer World, 16, 14.

Pine, Steven M. (1986). Reduction of test bias by adaptive testing, an unpublished manuscript. Richardson, Frank C. and Woolfolk, Robert L. Mathematics Anxiety. In I.G. Sarason (Ed.) Test Anxiety: Theory, Research, and Applications. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

Riggs, Matt L. (1988). The development of self-efficacy and outcome expectancy scales for general applications. A paper presented to Society of Industrial/Organizational Psychology conference.

Rosen, Larry D., Sears, Deborah C., and Weil, Michelle M. (1987a). A Manual for Administration and Scoring of Three Instruments, California State University, Dominguez Hills.

Rosen Larry D., Sears, Deborah C., and Weil, Michelle M. (1987b). Session VII Computerphobia, Behavioral Research, Methods, Instruments, and Computers, 19, 2, 167-179.

Rubin, Charles. (1983). Some people should be afraid of computers, Personal Computing, August, 55-58.

Sarason, I.G. and Mandler, G. (1952). Some correlates of test anxiety, Journal of Abnormal Psychology, 47, 810-817.

Sarason, I. G. (1957). Test anxiety, general anxiety, and intellectual performance, Journal of Cognitive Psychology, 21, (6), 485-490.

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

Sarason, I.G. (1961). Test anxiety and the intellectual performance of college students, Journal of Educational Psychology, 52, (4), 201-206.

Sarason, I. G. (1963). Test anxiety and intellectual performance, Journal of Abnormal Psychology, 66, 1, 73-75.

Sarason, I.G. Experimental approaches to test anxiety: Attention and the uses of information. In C.D. Spielberger (Ed.), Anxiety: Current trends in theory and research, (Vol.2). New York: Academic, 1972.

Sarason, I.G. (1973). Test anxiety and cognitive modeling, Journal of Personality and Social Psychology, 28, 58-61.

Sarason, I.G. The Test Anxiety Scale: Concept and research. In C.D. Spielberger and I. G. Sarason (Eds.), Stress and Anxiety, (Vol. 5). Washington: Hemisphere, 1978.

Sarason, I.G. Test Anxiety: Theory, Research, and Applications. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

Schmidt, F.L. and Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization, Journal of Applied Psychology, 62, (5), 529-540.

Schmidt, F.L. and Hunter, J.E. (1981). Employment testing: Old theories and new research findings, American Psychologist, 36, 1128-1137.

Sieber, Joan E. Define Test Anxiety: Problems and Approaches. In I.G. Sarason (Ed.) Test Anxiety: Theory, Research, and Applications. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

Skinner, H.A. & Allen, B.A. (1983). Does the computer make a difference? Computerized versus face-to-face self-report assessment of alcohol, drug, and tobacco use. Journal of Consulting and Clinical Psychology, 51, 267-275.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the Validation and Use of Personnel Selection Procedures. (Third Edition) College Park, MD: Author.

Spielberger, C.D. The effects of anxiety on complex learning and academic achievement. In C.D. Spielberger (Ed.), Anxiety and Behavior. New York: Academic Press, 1966.

Spielberger, C.D. Gorsuch, R.L., and Lushene, R.E. Manual for the State-Trait Anxiety Inventory, Palo Alto, Calif: Consulting Psychological Press, 1970.

Spielberger, C.D., and Sarason, I.G. Stress and Anxiety, (Vol.5) Washington: Hemisphere Publishing Corporation, 1966.

Spielberger, C.D., Gonzalez, H.P., Taylor, C.J., Algaze, B., Anton, W.D. Examination stress and test anxiety. In C.D. Spielberger and I.G. Sarason, Stress and Anxiety, (Vol. 5), Washington: Hemisphere, 1978.

Urry, Vern W., Tailored testing: A successful application of latent trait theory, Journal of Educational Measurement, 14, 2, Summer 1977, pp. 181-196.

Vale, C.D. and Weiss, D.J. (1978). The stratified adaptive ability test as a tool for personnel selection and placement. TIMS Studies in the Management Sciences, 8, 135-151.

Wang, Alvin Y. and Richarde, R. Stephen. (1988). Global versus task-specific measures of self-efficacy, The Psychological Record, 38, 533-541.

Weinberg, S. (1982). Coping with cyberphobia, Mademoiselle, August, 12.

Weiss, David J. (1982). Improving measurement quality and efficiency with adaptive testing, Applied Psychological Measurement, 6, 4, Fall, 473-492.

Weiss, David J. New Horizons in Testing. New York: Academic Press, 1983.

Wine, Jeri D. Cognitive-attentional theory of test anxiety. In I.G. Sarason (Ed.) Test Anxiety: Theory, Research, New York: Academic Press, 1978.

Wine, Jeri D. (1971) Test anxiety and direction of attention, Psychological Bulletin, 76, 92-104.

Wood, Robert E. and Locke, Edwin A. (1987). The relation of self-efficacy and grade goals to academic performance, Educational and Psychological Measurement, 47, 1013-1025.

Wright, B.D. (1977). Solving measurement problems with the Rasch model, Journal of Educational Measurement, 14, 97-116.

Wright, B.D. and Douglas, G.A. (1977). Conditional versus unconditional procedures for sample-free analysis, Educational and Psychological Measurement, 37, 573-586.

Wrightsman, Jr., Lawrence S. (1962). The effects of anxiety, achievement motivation, and task importance upon performance on an intelligence test, Journal of Educational Psychology, 53, (3), 150-156.